**6:38**

**Analytics Vidhya 14:13**

**Reference table**

ek aisi table hoti hai jo doosri tables ke liye fixed data store karti hai, jaise country codes ya product categories.

copying a table

Data analytics mein copying a table ka matlab hai ek table ka same data aur structure wala duplicate banana.

**Sampling technique**

ek tarika hai jisme bade data ka ek chhota hissa le kar uspar analysis kiya jata hai.

Types :

**Simple random sampling**

ek sampling technique hai jisme har individual ya item population mein barabar chance ke sath randomly select hota hai.

**Systematic sampling**

ek tarika hai jisme data ya logon ko ek fix gap (jaise har 5th ya 10th) ke saath chuna jata hai.

**Cluster sampling**

ek tarika hai jisme bade group ko chhoti-chhoti groups (clusters) mein baant kar, kuch groups ko randomly chuna jata hai aur unke sabhi logon ka analysis kiya jata hai.

**Stratified sampling**

ek tarika hai jisme population ko chhoti-chhoti groups (strata) mein baant kar har group se thoda-thoda sample liya jata hai, taaki data zyada accurate ho.

**Judgmental sampling**

mein researcher apni samajh se important samples chunta hai, bina kisi random process ke.

Hypothesis testing

ek statistical technique hai jo data ka analysis karke kisi assumption (null hypothesis) ko reject ya accept karne mein madad karti hai.

**Append vs Merge in Power BI -**

**What is the difference?**

Power BI me Append tables ko row-wise jodta hai (jaise union),

jabki Merge tables ko ek common column ke basis pe join karke naye columns add karta hai (jaise SQL join).

| Feature | Append | Merge |
| --- | --- | --- |
| Kya karta hai? | Ek table me doosri table ke rows add karta hai | Common column ke basis pe tables ko join karta hai |
| Kis tarah kaam karta hai? | SQL ke UNION ki tarah | SQL ke JOIN ki tarah |
| Structure par asar | Rows badhti hain | Columns badhte hain |
| Kab use karein? | Jab same structure wale data ko stack karna ho | Jab tables ko ek dusre se link karke additional info chahiye |
| Example | Multiple months ka sales data combine karna | Customer ID ke basis pe customer aur order history merge karna |

**How would you handle missing**

**dates in Power BI?**

Power BI me missing dates handle karne ke liye aap DAX ka use karke ek date table bana sakte hain (CALENDAR ya CALENDARAUTO), fact tables ke saath relationship set kar sakte hain, aur FULL OUTER JOIN ya CROSSJOIN ka use karke gaps fill kar sakte hain.

**Fill missing data**

Mean, mode, median se

**To calculate year-over-year (YoY) growth in Power BI using DAX, create this measure:**

YoY Growth =

VAR CurrentYear = SUM('Table'[Sales])

VAR LastYear = CALCULATE(SUM('Table'[Sales]), SAMEPERIODLASTYEAR('Table'[Date]))

RETURN DIVIDE(CurrentYear - LastYear, LastYear, 0)

**Explain Row-Level Security (RLS) and how you can implement it**

Row-Level Security (RLS) ek data access control technique hai jo users ko sirf unhi rows dekhne ya modify karne deti hai jo unke role ya attributes ke basis par allowed hoti hain, aur ise SQL policies (PostgreSQL, SQL Server), security predicates (SQL Server), ya ORMs ke application-level filtering se implement kiya ja sakta hai.

**Measures aur calculated columns**

Measures jab aap report dekhte hain tab har baar calculation karte hain,

lekin Calculated Columns sirf ek baar data load hone par calculate hote hain.

| **Feature** | **Measures** | **Calculated Columns** |
| --- | --- | --- |
| **Calculation** | Har baar report dekhne par hoti hai | Sirf ek baar data load hone par hoti hai |
| **Context** | Filter aur slicers ke saath change hoti hai | Hamesha same rehti hai |
| **Storage** | Memory nahi leti, sirf calculation hoti hai | Dataset ka hissa hoti hai, extra storage leti hai |
| **Use Case** | Aggregations jaise SUM, AVERAGE, COUNT | Naye derived columns banane ke liye |

**Outliers**

woh numbers hote hain jo kisi data mein baaki sabse alag dikhte hain.

Outliers woh numbers hote hain jo kisi dataset mein baaki values se bohot alag dikhte hain.

**How to Detect Outliers?**

Outliers detect karne ke liye Box plot, Z-score, IQR (Interquartile Range) jaise statistical methods ya phir Isolation Forest aur DBSCAN jaise machine learning techniques use kar sakte hain.

**How to treat Outliers?**

Outliers ko find karne ke liye kuch methods use karte hain, drop outliers

Assign new value

Cap outliers data

Try new transformations

phir unhe hata dete hain, adjust karte hain, ya limit karte hain, jo analysis pe depend karta hai.

**What are different data validation methods used by data analyst?**

**Data validation**

ek process hai jismein data ki accuracy, consistency, aur reliability ko check kiya jata hai, taaki galat ya incomplete data system mein na enter ho.

**Types of Data Validation**

FIELD LEVEL VALIDATION Field-level validation ka matlab hai, jab hum kisi form ya input field mein daale gaye data ko check karte hain, jaise ki email ya phone number, taaki wo sahi ho.

FORM LEVEL VALIDATION Form level validation ka matlab hai form bharne ke baad sabhi fields ko check karna, taake sab kuch sahi ho aur kuch galat na ho.

DATA SAVING VALIDATION Data saving validation ka matlab hai yeh check karna ki data sahi ho aur sirf zaroori data hi save ho, taaki space bacha rahe.

SEARCH CRITERIA VALIDATION Search criteria validation matlab search karte waqt jo terms ya filters diye jaate hain, unko check karna ki wo sahi hain ya nahi, taaki aapko sahi results mil sakein.

**Data Cleaning and preparation**



****Data analysis karne ke liye kuch basic steps hote hain jo raw data ko useful insights me convert karne me madad karte hain. Yahaan ek structured approach hai:

### **1. Problem ya Objective Define karo**

* Sabse pehle samjho ki kis problem ka solution nikalna hai ya kis sawal ka answer chahiye.
* Stakeholders se discuss karo aur unke expectations samjho.
* Success criteria decide karo.

### **2. Data Collect karo**

* Data alag-alag sources se collect karo (databases, APIs, surveys, Excel files, etc.).
* Ensure karo ki data reliable aur relevant ho.

### **3. Data Clean karo (Data Preprocessing)**

* Missing values ko handle karo (fill, remove, etc.).
* Duplicate aur inconsistent records ko remove karo.
* Data formats ko standardize karo (date-time, categories, etc.).
* Outliers aur unwanted noise ko remove karo.

### **4. Data ko Explore karo (Exploratory Data Analysis - EDA)**

* Summary statistics nikalo (mean, median, mode, standard deviation).
* Data ko graphs aur charts se visualize karo (histogram, scatter plot, box plot).
* Patterns, trends, aur correlations identify karo.

### **5. Data Transform aur Prepare karo**

* Feature engineering karo (naye variables create karo, normalization, encoding, etc.).
* Dimensionality reduction agar zaroori ho (PCA, feature selection).
* Machine learning ke liye train-test split banao.

### **6. Statistical ya Machine Learning Techniques Apply karo**

* Hypothesis testing, regression, clustering, classification apply karo.
* Machine learning models train aur optimize karo.
* Model fine-tune karo best performance ke liye.

### **7. Results Interpret aur Validate karo**

* Model ke performance metrics check karo (accuracy, RMSE, F1-score, etc.).
* Ensure karo ki results business objectives ke sath align ho rahe hain.
* Overfitting avoid karne ke liye cross-validation karo.

### **8. Insights aur Findings Communicate karo**

* Dashboards, reports, aur data visualizations create karo.
* Stakeholders ko insights simple language me samjhao (data storytelling).
* Data-driven recommendations do.

### **9. Action lo aur Monitor karo**

* Insights ke basis pe decisions lo.
* Changes implement karne ke baad performance track karo.
* Naye data ke sath analysis refine karte raho.

**Data analysis**

ka matlab hai data ko samajhna, uska dhyan se analysis karna aur usse useful jaankari nikalna.

**Data mining**

ek process hai jisme bade datasets se patterns, trends aur useful information extract ki jati hai taaki behtar decisions liye ja sakein.

**Diff**

Data mining naye patterns aur insights dhoondhne par focus karta hai, jabki data analysis in patterns ka interpretation aur decision-making ke liye istemal karta hai.

Data mining naye patterns dhoondhta hai, aur data analysis un patterns ko samajhkar faisle lene mein madad karta hai.

**State different steps in data cleansing activities**

Data cleansing ka process bohot important hai taaki jo data aap analyze kar rahe ho, wo accurate aur quality ho. Yeh kuch common steps hain jo data cleansing mein follow kiye jaate hain:

1. Duplicates Hatana:
   * Data mein jo duplicate records hain unhe identify karke remove karo taaki analysis skew na ho.
2. Missing Data Handle Karna:
   * Data mein missing ya null values ko identify karo.
   * Decide karo ki unko remove karein, fill karein (impute), ya business requirement ke hisaab se rakhein.
3. Data Formats Ko Standardize Karna:
   * Data formats ko consistent banao, jaise date format, address format, etc.
4. Inconsistent Data Correct Karna:
   * Agar data inconsistent hai, jaise uppercase-lowercase, spelling mistakes, toh unhe correct karo.
5. Data Accuracy Validate Karna:
   * Data ki accuracy ko reliable sources ya predefined rules ke saath cross-check karo.
6. Outliers Detect Karna:
   * Outliers ko identify karo jo analysis ko distort kar sakte hain aur unhe remove ya adjust karo business logic ke hisaab se.
7. Irrelevant Data Hatana:
   * Un columns ya records ko identify karo jo analysis mein help nahi karte aur unhe remove karo.
8. Data Normalize Karna:
   * Numerical data ko uniform scale pe laane ke liye normalize karo, taaki different sources se data consistent ho.
9. Data Transformation:
   * Data ko transform karo agar zaroorat ho (jaise categorical variable ko numerical mein convert karna) taaki analysis mein help ho sake.
10. Data Integrity Issues Address Karna:
    * Agar data integrity violate ho rahi hai, jaise start date end date se baad ho, toh un issues ko resolve karo.

Yeh steps follow karke aap apna data clean kar sakte ho, jo analysis ke liye reliable aur accurate hoga.

**Explain different tools used for data analysis.**

Data analysis ke liye kai tools available hain jo data ko collect, process, aur analyze karne mein madad karte hain. Yeh kuch popular tools hain jo data analysis mein use kiye jaate hain:

1. Microsoft Excel

* Purpose: Small-scale data analysis aur visualization ke liye widely use hota hai.
* Features: Formulas, pivot tables, data filtering, aur basic charts banane ki facility.
* Best for: Quick analysis, reports create karna, aur simple data visualization.

2. Power BI

* Purpose: Microsoft ka ek business analytics tool jo interactive visualizations aur business intelligence ke liye use hota hai.
* Features: Easy data importing, transforming, aur dashboards aur reports create karne ki facility.
* Best for: Large datasets ke liye aur real-time reports banane ke liye.

3. Tableau

* Purpose: Data visualization aur business intelligence tool.
* Features: Drag-and-drop feature se interactive dashboards aur reports banane ka option.
* Best for: Data visualization, decision-makers ke liye insightful dashboards banane ke liye.

4. R

* Purpose: Statistical computing aur graphics ke liye programming language.
* Features: Advanced statistical analysis, machine learning algorithms aur visualization tools.
* Best for: Complex statistical analysis aur large datasets ke liye.

5. Python

* Purpose: General-purpose programming language jo data analysis ke liye libraries provide karta hai.
* Features: Pandas, NumPy, Matplotlib, Seaborn, aur Scikit-learn jaise libraries se data cleaning, analysis aur visualization ki jaati hai.
* Best for: Machine learning, data manipulation, aur repetitive tasks ko automate karne ke liye.

6. SQL (Structured Query Language)

* Purpose: Relational databases se data manage karne aur query karne ke liye language.
* Features: Data ko query, update, aur manage karne ki facility.
* Best for: Databases se data extract karke analysis karne ke liye.

**Explain data purging?**

Data purging ka matlab hai unnecessary, purani, ya redundant data ko database ya data system se permanently remove karna. Yeh process system ki performance ko optimize karne, storage cost ko reduce karne, aur data quality ko improve karne ke liye kiya jata hai. Isme aise records ya files ko delete karna hota hai jo ab current operations ya compliance ke liye necessary nahi hote.

Data purging ke kuch main reasons:

1. Storage space free karna: Purani ya unused data ko remove karne se disk space save hota hai, especially jab system mein bahut zyada data ho.
2. Performance improve karna: Jab unnecessary data ko hata diya jata hai, toh database aur systems zyada efficiently kaam karte hain.
3. Data quality maintain karna: Purani ya galat data ko purge karke hum ensure karte hain ki jo data bacha hai, wo accurate aur relevant ho analysis aur decision-making ke liye.
4. Compliance: Kuch industries mein regulations hote hain jo demand karte hain ki kuch types ka data ek certain period ke baad delete kar diya jaye, jaise privacy laws ke under personal information.

Example: Agar ek company customer records jo 10 saal purane hain, unhe purge kar deti hai, ya server logs ko purge kar deti hai jo already analyze ho chuke hain.

Purging karte waqt, dhyan rakhna zaroori hai taaki important data galti se delete na ho jaye.

**What will be common issues faced by data analyst?**

As a data analyst, yeh kuch common issues ho sakte hain jo tumhe face karne pad sakte hain:

1. Data Quality Issues: Kabhi-kabhi data incomplete, inaccurate ya inconsistent hota hai. Missing values, incorrect entries ya duplicates ke wajah se analysis reliable nahi hota.
2. Data Integration: Alag-alag sources se data ko combine karna mushkil ho sakta hai. Data ko compatible banane aur consistency ensure karne ki zarurat hoti hai.
3. Data Cleaning: Analysis se pehle data ko clean karna time-consuming ho sakta hai. Isme errors ko identify karna, irrelevant data ko remove karna, aur outliers ko handle karna padta hai.
4. Handling Big Data: Jab data bahut zyada ho, to usse process aur analyze karna difficult ho sakta hai. Large datasets ko efficiently manage karne ke liye proper tools ki zarurat hoti hai.
5. Complexity of Analysis: Kabhi-kabhi large aur complex datasets ko analyze karna tricky hota hai. Har type ke analysis ke liye appropriate statistical methods aur algorithms samajhna zaruri hai.
6. Communication of Results: Findings ko clear aur understandable tarike se non-technical logon ko present karna mushkil ho sakta hai. Complex data ko simplify karna, bina accuracy lose kiye, ek challenge ho sakta hai.
7. Tool and Software Limitations: Power BI, Excel ya Python jaise tools mein kabhi-kabhi limitations hoti hain, especially jab complex datasets ho ya advanced analysis ki zarurat ho.
8. Keeping Up with New Tools and Techniques: Data analysis ka field bohot tezi se evolve ho raha hai. Naye tools aur techniques ke saath updated rehna thoda challenging ho sakta hai.
9. Data Security and Privacy: Sensitive data ke saath kaam karte waqt data security aur privacy regulations ka dhyan rakhna bohot zaruri hai.
10. Time Constraints: Data analysis kaam ko deadlines ke andar complete karna padta hai, jo kabhi-kabhi rushed analysis ka sabab ban sakta hai.

**Explain different usages of data mining.**

Data mining ek process hai jisme hum large data sets se patterns, correlations aur useful information nikalte hain. Iska use kai different fields mein hota hai. Kuch common usages yeh hain:

1. Customer Relationship Management (CRM):
   * Businesses ko customer ke behavior, preferences aur purchasing patterns samajhne mein help milti hai. Isse wo apni marketing strategies personalize kar sakte hain aur product recommendations ko improve kar sakte hain.
2. Fraud Detection:
   * Financial institutions aur online retailers data mining ka use fraud detect karne ke liye karte hain. Jab koi unusual pattern ya anomaly transaction data mein hoti hai, toh fraud ka pata chal jata hai.
3. Market Basket Analysis:
   * Retail mein yeh technique use hoti hai jisme dekha jata hai ki kaunse items customers ek sath purchase karte hain. Jaise agar bread aur butter hamesha ek sath khareede jate hain, toh store unhe bundle kar sakta hai ya promotion design kar sakta hai.
4. Healthcare:
   * Healthcare mein, data mining se disease outbreaks predict karna, patient outcomes samajhna aur risk factors identify karne mein help milti hai. Yeh drug discovery aur hospital operations ko bhi optimize karta hai.
5. Stock Market Analysis:
   * Data mining stock market trends predict karne ke liye use hoti hai, jisme historical data jaise price fluctuations, trading volumes aur economic indicators analyze kiye jate hain. Isse investors ko better decisions lene mein madad milti hai.
6. Text Mining:
   * Yeh technique large volumes of unstructured text data jaise customer reviews, social media posts aur news articles ko analyze karne ke liye use hoti hai. Isse sentiment analysis, emerging trends aur customer feedback analysis hota hai.
7. Quality Control:
   * Manufacturing industries mein data mining se defects identify karna, processes improve karna aur production schedules optimize karna asaan ho jata hai. Yeh sensor data aur production lines se milti hai.
8. Predictive Maintenance:
   * Industries jaise aviation aur manufacturing mein data mining se machines ya equipment ki failure prediction hoti hai, jisse maintenance time pe ho sakta hai aur downtime reduce hota hai.
9. Education:
   * Educational institutions mein data mining se student performance track kiya jata hai, at-risk students ko identify kiya jata hai aur learning paths optimize kiye jate hain. Yeh personalized learning experiences banane mein help karta hai.
10. E-commerce Personalization:
    * Online shopping platforms data mining ka use user preferences, browsing behavior aur purchase history ko samajhne ke liye karte hain. Isse wo personalized product recommendations de sakte hain.
11. Social Network Analysis:
    * Social media mein data mining user interactions, connections aur content sharing patterns ko analyze karne ke liye use hoti hai. Isse influencers identify kiye jate hain aur targeted ads ko improve kiya jata hai.

**Tell different industries where data analysis is frequently used?**

Data analysis ka use kai industries mein hota hai. Yeh kuch major industries hain jahan data analysis kaafi important hai:

1. Healthcare:
   * Patient data analyze karke treatment plans improve karna, disease outbreaks predict karna, aur hospital management optimize karna.
   * Drug development aur clinical trials mein bhi data analysis ka use hota hai.
2. Retail:
   * Customer behavior samajhna, sales patterns aur inventory manage karna.
   * Marketing strategies ko personalize karna aur promotions ko optimize karna.
3. Finance and Banking:
   * Fraud detect karna, stock market trends predict karna, aur risk management.
   * Financial decisions ko improve karne ke liye customer segmentation.
4. E-commerce:
   * Customer data analyze karke personalized recommendations dena, demand forecasting, aur supply chain ko optimize karna.
5. Education:
   * Student performance data analyze karke learning ko personalize karna aur outcomes improve karna.
   * Resources optimize karna aur learning behaviors ke trends samajhna.
6. Telecommunications:
   * Network performance monitor karna, customer churn predict karna, aur pricing strategies ko optimize karna.
7. Manufacturing:
   * Predictive maintenance, production efficiency improve karna, aur operational costs ko reduce karna.
8. Transportation and Logistics:
   * Route planning optimize karna, delivery times improve karna, aur traffic patterns analyze karna.
   * Vehicles aur infrastructure ki predictive maintenance.
9. Marketing and Advertising:
   * Customer demographics aur ad performance analyze karna, aur campaign effectiveness improve karna.
10. Energy:
    * Demand aur supply predict karna, energy production optimize karna, aur wastage reduce karna.
    * Energy consumption patterns ko monitor karke sustainability improve karna.
11. Sports:
    * Player performance aur game strategy ko optimize karna.
    * Injury prediction aur prevention through data analysis.
12. Government and Public Sector:
    * Census data, public health data, aur traffic patterns analyze karke policymaking improve karna.
    * Resources allocate karna aur budgeting mein help karna.

**Explain different skills where the data analysis needs to be used.**

Data analysis is an essential skill that spans various industries and fields. Here are some key areas where data analysis is crucial:

1. Business Intelligence (BI):
   * Skills Required: Data visualization, reporting, forecasting, and performance tracking.
   * How It's Used: Analyzing sales, customer data, and market trends to help businesses make strategic decisions and improve operational efficiency.
2. Healthcare Analytics:
   * Skills Required: Statistical analysis, pattern recognition, knowledge of medical terminology, and data visualization.
   * How It's Used: Analyzing patient data, treatment outcomes, and hospital performance to improve healthcare services and predict future trends in patient care.
3. Finance and Accounting:
   * Skills Required: Financial modeling, data forecasting, risk analysis, and Excel proficiency.
   * How It's Used: Analyzing market data, investments, and financial reports to make informed decisions about budgets, investments, and economic strategies.
4. Marketing and Customer Insights:
   * Skills Required: Customer segmentation, trend analysis, A/B testing, and social media analytics.
   * How It's Used: Understanding consumer behavior, evaluating marketing campaigns, and optimizing customer engagement strategies to enhance brand loyalty and sales.
5. Supply Chain and Operations:
   * Skills Required: Logistic regression, demand forecasting, optimization models.
   * How It's Used: Streamlining operations, optimizing inventory management, and improving product distribution by analyzing patterns in logistics, suppliers, and demand.
6. Sports Analytics:
   * Skills Required: Statistical modeling, performance analysis, and machine learning.
   * How It's Used: Analyzing player performance, team dynamics, and game strategies to enhance athletic performance and optimize training.
7. Government and Public Policy:
   * Skills Required: Data interpretation, statistical analysis, and public policy knowledge.
   * How It's Used: Analyzing demographic, economic, and social data to inform policies, track government performance, and plan for future societal needs.
8. Education and Academic Research:
   * Skills Required: Statistical methods, survey analysis, and hypothesis testing.
   * How It's Used: Analyzing student performance, conducting research studies, and evaluating educational policies to improve learning outcomes.
9. Human Resources (HR):
   * Skills Required: Employee performance analysis, recruitment data analysis, and retention strategies.
   * How It's Used: Analyzing employee data to improve recruitment strategies, enhance workplace productivity, and reduce turnover rates.
10. Technology and Software Development:
    * Skills Required: Data cleaning, coding (Python, R), algorithm development.
    * How It's Used: Analyzing usage patterns, bug reports, and customer feedback to improve software functionality and optimize the user experience.

### **Daily Tasks of a Data Analyst (in Simple English)**

1. **Collecting Data**
   * Gather data from different sources like databases, APIs, or CSV files.
   * Check if the data is accurate and complete.
2. **Cleaning and Preparing Data**
   * Fix errors like duplicate or missing values.
   * Format the data properly for analysis.
3. **Exploring and Analyzing Data**
   * Look for patterns, trends, and relationships in the data.
   * Use tools like Excel, SQL, Python, or R.
4. **Creating Reports and Visualizations**
   * Make charts, graphs, and dashboards using Power BI, Tableau, or Excel.
   * Present insights in a clear way to stakeholders.
5. **Working with Teams**
   * Talk to managers or clients to understand their needs.
   * Collaborate with teams like marketing, sales, and product.
6. **Writing SQL Queries**
   * Write SQL queries to get the required data from databases.
   * Automate queries for repeated tasks.
7. **Solving Business Problems**
   * Understand business challenges and find solutions using data.
   * Provide recommendations based on insights.
8. **Maintaining Documentation**
   * Keep records of data sources, methods, and processes.
   * Document work for future reference.
9. **Learning New Tools and Trends**
   * Stay updated on the latest tools and techniques.
   * Improve skills in Python, SQL, and data visualization.
10. **Monitoring and Fixing Issues**

* Ensure reports and dashboards are accurate.
* Fix errors in data pipelines or visualizations.

### **Commonly Used Tools**

* **Data Query & Cleaning:** SQL, Excel, Python, R
* **Visualization:** Power BI, Tableau, Excel
* **Collaboration:** Jira, Slack, Trello
* **Analytics Platforms:** Google Analytics, Microsoft Azure, AWS

### **Different Phases of Data Analytics (Simple English)**

Data analytics is a step-by-step process that turns raw data into useful insights. It has several phases:

### **1. Define Objective (Set Goal)**

**Purpose:** Understand the problem or goal.  
 **What to do:**

* Identify the business problem or question.
* Set key performance indicators (KPIs) to measure success.

### **2. Data Collection (Gather Data)**

**Purpose:** Collect data from useful sources.  
 **What to do:**

* Get data from databases, APIs, surveys, or other sources.
* Ensure the data is accurate and reliable.

### **3. Data Cleaning and Preparation**

**Purpose:** Make raw data usable by fixing errors.  
 **What to do:**

* Handle missing data and remove duplicates.
* Standardize formats and correct mistakes.
* Convert data into the required structure.

### **4. Data Exploration and Analysis**

**Purpose:** Understand data and find patterns.  
 **What to do:**

* Use charts and graphs for visualization.
* Identify trends, relationships, and outliers.
* Perform Exploratory Data Analysis (EDA).

### **5. Data Modeling**

**Purpose:** Create models for predictions and insights.  
 **What to do:**

* Use statistical or machine learning methods.
* Test and optimize models.
* Select the best model for analysis.

### **6. Data Visualization and Communication**

**Purpose:** Present insights in a clear way.  
 **What to do:**

* Create charts, dashboards, and reports.
* Share insights in an easy-to-understand format.
* Highlight key findings and action points.

### **7. Decision-Making and Implementation**

**Purpose:** Use insights to make smart decisions.  
 **What to do:**

* Recommend strategies based on findings.
* Take actions and track results.

### **8. Monitoring and Evaluation**

**Purpose:** Check the impact of decisions and improve them.  
 **What to do:**

* Track performance using KPIs.
* Implement feedback loops for continuous improvement.

This process helps businesses make data-driven decisions efficiently! 🚀

**What is Data Cleansing?** Data Cleansing (or data cleaning) is the process of identifying and fixing errors, inaccuracies, or inconsistencies in a dataset. The goal is to make the data clean, accurate, and ready for analysis.

### **Steps of Data Cleansing:**

1. **Identifying Errors:**
   * Missing values (e.g., empty cells)
   * Duplicate entries
   * Outliers or unusual data points
   * Inconsistent formats (e.g., different date or text formats)
   * Incorrect data (e.g., spelling mistakes or wrong values)
2. **Fixing Errors:**
   * Filling missing values (using mean, median, or prediction techniques)
   * Removing duplicate data
   * Standardizing data formats (e.g., converting all text to lowercase)
   * Correcting incorrect values
   * Handling outliers (removing or adjusting them)
3. **Validation and Verification:**
   * Checking the corrected data to ensure accuracy
   * Ensuring data follows business rules or required formats

### **Why is Data Cleansing Important?**

* **Improves Data Quality:** Ensures accurate and reliable data.
* **Better Decision-Making:** Clean data leads to better insights and smarter decisions.
* **Increases Efficiency:** Reduces errors and saves time.
* **Ensures Compliance:** Helps meet regulatory standards.

### **Data Cleansing Tools:**

* **Excel:** Basic cleaning tasks like removing duplicates and finding missing values.
* **Python:** Uses the Pandas library for advanced data cleaning.
* **Power BI:** Uses Power Query to transform data.
* **SQL:** Cleans data at the database level.