Contents

Overview of Data Analytics

Data analytics involves examining datasets to draw conclusions, identify patterns, and make informed decisions. It encompasses techniques from statistics, computer science, and business intelligence.

Types of Data Analytics:

- Descriptive Analytics: What happened? Summarizes historical data.
- Diagnostic Analytics: Why did it happen? Analyzes the cause of trends or anomalies.
- Predictive Analytics: What will happen? Uses models to forecast future trends.
- Prescriptive Analytics: What should we do? Recommends actions based on insights.

Applications:

- ✓ Healthcare: Predict patient health outcomes.
- ✓ Finance: Detect fraudulent transactions.
- ✓ Retail: Personalize customer experiences.

Example: A retail company uses analytics to identify that customers aged 25-34 frequently purchase outdoor gear. This insight drives targeted marketing campaigns.

Industry Analysis/ Market/ Players

Data Analytics Market:

- Estimated to grow rapidly due to increasing data generation.
- Industries leveraging analytics include e-commerce, healthcare, finance, logistics, and entertainment.

Major Players:

- Google (BigQuery): Cloud-based data warehouse solutions.
- Microsoft (Power BI, Azure Analytics): Tools for visualization and AI integration.
- IBM (Watson Analytics): AI-driven analytics solutions.
- Tableau (Salesforce): Intuitive dashboards and visualizations.
- Snowflake: Cloud-based data warehousing.

Example: A healthcare analytics firm like Optum uses AI to improve patient care and reduce costs.

Importance of Data

- ✓ Decision-Making: Reliable data drives better business decisions.
- ✓ Competitive Edge: Data insights give companies an advantage over competitors.
- ✓ Operational Efficiency: Optimizes resources and reduces waste.
- ✓ Customer Insights: Improves customer satisfaction and loyalty.

Example: Netflix uses data on viewing habits to recommend shows, resulting in increased viewer engagement.

Principles of Data

- ✓ Accuracy: Data must be error-free and reliable.
- ✓ Consistency: Uniform data ensures coherence across systems.
- ✓ Timeliness: Data should be available when needed.
- ✓ Completeness: Missing data can lead to flawed insights.
- ✓ Relevance: Data must serve the intended purpose.

Example: In predictive analytics, irrelevant variables can lead to overfitting in models, reducing their real-world applicability.

A Review on Descriptive Statistics

Definition: Descriptive statistics summarize and describe the characteristics of a dataset.

Central Tendency:

- Mean: Average value.
- Median: Middle value in sorted data.
- Mode: Most frequent value.

Dispersion:

- Range: Difference between the maximum and minimum.
- Variance: Spread of data around the mean.
- Standard Deviation: Square root of variance.

Shape of Distribution:

- Skewness: Measures symmetry.
- Kurtosis: Measures peakedness.
- Standard Deviation: Calculated using variance formula.

Descriptive statistics are fundamental for understanding data distributions before applying advanced analytics techniques.

From Business Problems to Data Mining Tasks

The journey from a business problem to a data mining task involves understanding the problem, identifying relevant data, and framing it into a task that can leverage data mining techniques.

Steps to Translate Business Problems:

1. Define the Problem: Understand the objectives.
    - Example: A retail store wants to reduce customer churn.

2. Identify the Data: Collect relevant data.
    - Example: Purchase history, customer demographics, service feedback.

3. Frame as Data Mining Tasks: Match the business need with data mining methods.
    - Predictive Task: Predict which customers are likely to churn.
    - Descriptive Task: Identify patterns in customer purchasing behavior.
    - Select the Technique: Choose algorithms such as classification, clustering, or regression.

The Types of Data

1. Structured Data: Organized into rows and columns (e.g., databases, spreadsheets).
Examples: Sales records, financial transactions.

2. Unstructured Data: Doesn't fit neatly into tables (e.g., text, images, videos).
Examples: Social media posts, medical images.

3. Semi-Structured Data: Contains elements of both (e.g., JSON, XML files).
Examples: Web logs, IoT data.

4. Temporal and Spatial Data:

- Temporal: Time-stamped data (e.g., stock prices).
- Spatial: Location-based data (e.g., GPS coordinates).

Sampling Techniques

Sampling is crucial when working with large datasets or when computational efficiency is a concern.

1. **Random Sampling: Every data point has an equal chance of being selected.**
   Example: Drawing 100 customer IDs randomly from a database.
2. **Stratified Sampling: Ensures each subgroup (strata) is proportionally represented.**
   Example: Sampling 20% of customers from each region.
3. **Systematic Sampling: Selects every *n*th data point.**
   Example: Choosing every 10th transaction.
4. **Cluster Sampling: Divides data into clusters and selects entire clusters.**
   Example: Surveying all customers in randomly chosen stores.

5. **Oversampling and Undersampling: Used to balance datasets in cases of class imbalance (e.g., fraud detection).**

Sampling is useful when you want to work with a smaller, representative subset of your data. Here's a step-by-step workflow for performing sampling in both Excel and Power BI.

Sampling in Excel

1. Random Sampling using RAND()
   (Data is in range A2:F101 (100 rows))

   Steps:
   a. In a new column (e.g., Column G), enter:

   ```
   =RAND()
   ```

   *This generates a random number between 0 and 1.

   b. Copy the formula down for all rows.
   c. Select the entire dataset including the RAND column.
   d. Go to Data tab → Click Sort → Sort by the RAND column (smallest to largest or vice versa).
   e. Pick the top n rows (e.g., first 10 rows for a 10% sample).

2. Using Data Analysis Toolpak (Simple Random Sampling)
   If enabled:
   a. Go to Data > Data Analysis > Sampling.
   b. Choose Input Range (include column headers if checked).
   c. Choose Random sampling.
   d. Enter the sample size.
   e. Choose Output Range or new worksheet.
   f. Click OK.

Sampling in Power BI

1. Random Sampling with Power Query
   Steps:
   a. Load your dataset into Power BI.
   b. Click Transform Data → opens Power Query Editor.
   c. Add a new column:
   d. Go to Add Column > Custom Column, and enter:

   ```
   Number.RandomBetween(1, 100)
   ```

   or

   ```
   Number.Random()
   ```

   *This generates a random number per row.

   e. Sort the column (ascending/descending).
   f. Use Keep Rows > Keep Top Rows (e.g., 10 or 20 rows).

2. Using DAX for Random Sampling (Report View)
   (Example, table is named Sales)

   Add a calculated column:

```
RandomValue = RAND()
```

   Create a new table with top N samples:

```
SampledSales = TOPN(10, Sales, Sales[RandomValue])
```

   *You can now use SampledSales as a new table in visuals.

3. Sampling in Power BI via Filters
   - If you don't need a perfectly random sample:
     a. Use slicers or filters to take a subset:
        i. Filter by Region, Product, Date Range, etc.
        ii. Manually select a few values

The Data Mining Process

Data mining is the process of:

- Discovering patterns
- Extracting useful insights
- Predicting future trends
- Using statistical and computational methods

What You Can Do in Excel (Without Add-ins)

- Descriptive Statistics → Explore basic data properties
- Correlation & Regression → Identify relationships
- Clustering / Segmentation → Use PivotTables and filters
- Classification / Forecasting → Limited, but some predictive modeling possible
- Association → Identify common co-occurrence patterns

Sample Dataset: Customer Purchases

Customer ID	Age	Region	Product	Category	Quantity	Total Spent
C001	25	East	Laptop	Electronics	1	1000
C002	30	West	Headphones	Electronics	2	200
C003	22	East	T-Shirt	Clothing	3	90
C004	45	Central	Refrigerator	Appliances	1	800
C005	35	West	Jacket	Clothing	1	120
C006	29	East	Laptop	Electronics	1	950
C007	38	Central	Microwave	Appliances	1	300
C008	26	West	T-Shirt	Clothing	2	60

Data Mining Workflow in Excel

1. Load and Prepare Data
   - Paste or import the dataset into Excel.
     a. Format it as a Table: Insert > Table
     b. Clean the data:
        i. Remove duplicates
        ii. Fill or delete missing values
        iii. Convert data types (Age = Number, Total Spent = Currency)

2. Descriptive Analysis (Profiling)
   Use formulas:

Average Age	=AVERAGE(B2:B9)
Total Revenue	=SUM(G2:G9)
Average Spend per Customer	=AVERAGE(G2:G9)
Most common product	=MODE(C2:C9) (if numeric) or use PivotTable

3. Segment Data (Clustering-like)
   Using PivotTables:
   - Insert PivotTable
     a. Rows → Region
     b. Values → Sum of Total Spent, Average Age

   - This helps identify:
     a. Top performing regions
     b. Age trends by region

   *You've now clustered by region.

4. Association Mining (Basket Analysis)

Count co-occurrence:
Use COUNTIFS() or PivotTables to find how often products are purchased together.

Example:
How many people aged 25–30 bought Electronics?

```
=COUNTIFS(B2:B9,">=25",B2:B9,"<=30",E2:E9,"Electronics")
```

5. Predictive Modeling (Regression)
   Go to Data > Data Analysis > Regression (Enable Analysis Toolpak if not visible)

   Input:
   - Y Range (Dependent Variable): Total Spent
   - X Range (Independent Variables): Age or Quantity

   Click OK
   *This shows if Age or Quantity can predict Total Spent.

6. Correlation Matrix
   Go to Data > Data Analysis > Correlation
   - Select numerical columns: Age, Quantity, Total Spent
   - Result: How strongly each variable relates to another

(Optional) Install Excel Data Mining Add-ins (for SQL Server)
If you're working with SQL Server, you can install:
- Microsoft Data Mining Add-ins for Excel

Gives access to:
- Classification
- Regression Trees
- Forecasting
- Key Influencer analysis

Comparing Techniques

Feature	EDA	Data Profiling	Data Mining	Data Analytics
Purpose	Explore and understand data	Assess data quality and structure	Discover patterns and predictions	Gain insights to support decisions
Focus	Patterns, anomalies, trends	Data validity, completeness	Hidden insights, patterns	Trends, KPIs, performance metrics
User	Analysts, Data Scientists	Data Engineers, QA	Data Scientists, ML Engineers	Business Analysts, Execs
Typical Tools	Excel, Python, R, Power BI	SQL, Power Query, ETL Tools	Python (ML), R, Weka, SSAS	Power BI, Tableau, Excel, SQL
Outputs	Visuals, stats, distributions	Null counts, data ranges	Clusters, models, rules	Dashboards, reports, summaries

The “Big Data”

Definition: Refers to datasets that are too large or complex to be processed using traditional methods.

Characteristics (5 V’s):

1. Volume: Massive amounts of data.
   Example: Petabytes of social media posts.

2. Velocity: The speed of data generation.
   Example: Real-time financial transactions.

3. Variety: Different data types.
   Example: Text, images, videos, and structured tables.

4. Veracity: Data accuracy and quality.
   Challenge: Cleaning noisy or inconsistent data.

5. Value: Extracting insights to drive decisions.
   Example: Personalized recommendations on e-commerce sites.

Big Data Technologies:

- Hadoop: Distributed data storage and processing.
- Spark: In-memory processing for faster computation.
- NoSQL Databases: MongoDB, Cassandra for semi-structured data.
- Cloud Platforms: AWS, Azure, GCP for scalable data solutions.

Applications:

- Healthcare: Genomic data analysis.
- E-commerce: Predicting customer behavior.
- Finance: Fraud detection using real-time transaction analysis.

Example Use Case for Integration: A telecom company wants to reduce churn.

- Big Data Role: Analyze millions of call logs, complaints, and service usage data.
- Data Mining: Cluster customers by usage patterns and apply predictive models to identify churn risks.

Handling Big Data in Excel

Excel is not designed for big data, but you can optimize it:

Excel Limitations:

- Max rows per sheet: 1,048,576
- Slows down with large file sizes (>100MB)
- Limited memory management (relies on RAM)

Strategies to Handle Big Data in Excel:

1. Use Power Query (Get & Transform)
    - Load only needed columns and rows
    - Apply filters early to reduce data volume
    - Use Power Query's "Enable Fast Load" feature

2. Use Excel Data Model (Power Pivot)
    - Load data into the Data Model, not the worksheet
    - Create relationships and measures using DAX
    - Excel compresses data more efficiently in the Data Model

3. Avoid Volatile Formulas
    - Use =SUMIFS() instead of =ARRAYFORMULA() or volatile functions like OFFSET() or INDIRECT()

4. Use External Connections
    - Connect Excel to SQL Server, Azure, or Power BI datasets to query only needed data
    - Use Data > Get Data > From Database or Web/ODBC

5. Split or Archive Data
    - Keep raw data in CSV/Text files and use queries to retrieve parts
    - Archive older records

Handling Big Data in Power BI

Power BI is much better suited for big data analytics with robust connectors and data modeling capabilities.

Key Features for Big Data Handling:

1. Use Power BI Dataflows (in Power BI Service)
Preprocess and transform data in the cloud
Reuse across reports and teams

2. DirectQuery or Live Connections
    - Don't import full dataset
    - Connect live to large databases like:
        - Azure Synapse
        - SQL Server
        - Snowflake
        - Databricks

3. Aggregation Tables
    - Build pre-aggregated summary tables in the model
    - Faster than querying raw, detailed data

4. Incremental Refresh
    - Only load new or changed data instead of reloading full datasets
    - Set up in Power BI Pro or Premium: Model > Table > Incremental Refresh

5. Optimize DAX and Model Design
    - Avoid complex calculated columns (use Power Query instead)
    - Use star schema modeling
    - Limit column cardinality (fewer unique values = faster performance)

Example: Large Dataset Strategy in Power BI

1. Connect to a SQL Server using DirectQuery

    Use Power Query to:
    - Filter out unnecessary columns
    - Normalize and flatten nested data

    In the data model:
    - Use summary/aggregation tables for visuals
    - Enable incremental refresh

2. Build visuals using measures, not calculated columns
3. Publish to Power BI Service and schedule data refresh

Getting Started with Analysis

Steps to Start Analysis:

1. Understand the Objective: Clearly define what you're trying to achieve.
   Example: Analyze monthly sales to identify declining trends.
2. Collect Data: Gather data from reliable sources.
   Example: Sales records, customer demographics, and product categories.
3. Clean the Data:
    - Handle missing values.
    - Remove duplicates.
    - Correct inconsistent data formats.
4. Perform Exploratory Data Analysis (EDA):
5. Summarize data with statistics like mean, median, or mode.
6. Visualize trends using charts.

Choose the Right Tools:

- Excel: For basic analysis and visualization.
- Minitab: For statistical tests like hypothesis testing or regression.

Organizing Data for Analysis

Best Practices:

1. Create a Data Dictionary: Define what each column represents. Example: "Sales" (USD) = Monthly revenue from products sold.
2. Structure Data into Tables:
   - Use rows for observations (e.g., customers).
   - Use columns for attributes (e.g., age, purchase amount).
3. Normalize Data: Reduce redundancy by creating related tables.
4. Use Descriptive Column Names: Avoid ambiguous labels. Example: Use “Customer_ID” instead of “ID.”
5. Sort and Filter Data: Arrange by relevant metrics, such as date or region.

Example:
An e-commerce dataset might be organized as:

Customer_ID	Date	Product	Quantity	Total_Sale
101	2024-11-01	Laptop	1	1200
102	2024-11-02	Phone	2	1500

The Language of Data

Understanding and using the "language of data" is critical for clear communication.

Key Terms:

- Variable: A characteristic being measured (e.g., Sales).
- Distribution: Spread of data values.
- Correlation: Relationship between variables.
- Outliers: Data points significantly different from others.

Example:
Describing data to stakeholders: "Our analysis shows a strong positive correlation of 0.85 between advertising expenditure and monthly sales, indicating higher spending results in increased revenue."

Communication Techniques with Data

Effective data communication involves translating findings into actionable insights.

1. Choose the Right Visualization:
   - Trends: Line charts.
   - Comparisons: Bar charts.
   - Relationships: Scatter plots.
   - Proportions: Pie charts.

2. Use Storytelling: Frame your findings as a narrative.
Example: "Sales grew by 15% this quarter due to targeted marketing."

3. Simplify Complex Metrics: Avoid jargon; use analogies if needed.
Example: "This algorithm is like a teacher identifying top-performing students."

4. Highlight Key Insights: Use callouts, annotations, or bold colors.
Example: Emphasize a 20% increase in revenue.

5. Tailor to the Audience:
- Executives: Focus on business impact.
- Analysts: Include detailed metrics.

Performing Exploratory Data Analytics (EDA)

Performing Exploratory Data Analysis (EDA) in Excel and Power BI involves examining datasets to summarize their main characteristics, often with visual methods.

EDA in Excel

1. Load and Clean Your Data
Use Power Query to load your dataset (from CSV, Excel file, SQL, etc.)

Use Power Query for:
- Removing duplicates
- Handling missing values
- Changing data types
- Filtering rows
- Creating calculated columns

2. Descriptive Statistics
Use Excel functions or Data Analysis Toolpak (enable via File > Options > Add-ins):
- Mean, Median, Mode → =AVERAGE(), =MEDIAN(), =MODE.SNGL()
- Standard Deviation, Variance → =STDEV.S(), =VAR.S()
- Min/Max → =MIN(), =MAX()
- Count, Unique Count → =COUNT(), =COUNTA(), =UNIQUE()

3. Visualizations
Use charts to explore patterns:
- Histograms (via Data Analysis Toolpak or manually using bins)
- Box Plots (using custom plots or templates)
- Scatter Plots
- Pivot Charts

- Bar/Column Charts

4. Pivot Tables
Use PivotTables to:

- Aggregate data
- Analyze categorical features
- Drill into distributions and groupings

Sample walkthrough of performing EDA in Excel using Power Query and Excel tools

Sample Dataset: Sales Data

Date	Region	Product	Category	Sales	Quantity
2024-01-01	East	Pencil	Stationery	120	10
2024-01-02	West	Binder	Stationery	250	5
2024-01-03	Central	Desk	Furniture	300	2
2024-01-04	East	Pen	Stationery	150	15
2024-01-05	West	Chair	Furniture	500	4
2024-01-06	Central	Pencil	Stationery	100	8
2024-01-07	East	Binder	Stationery	200	6
2024-01-08	West	Desk	Furniture	450	3

Step-by-Step EDA in Excel

1. Load the Data into Power Query
   a. Select the range (include headers).
   b. Go to the Data tab → Click From Table/Range.
   c. It prompts to create a table – click OK.
   d. Power Query Editor will open with your data.

2. Explore the Data in Power Query

In Power Query Editor:
You can enable:

- ✓ Column quality → shows errors, blanks, valid values.
- ✓ Column distribution → shows unique values.
- ✓ Column profile → gives stats (min, max, avg, etc.)

Clean the data (if needed):

- ✓ Rename columns for clarity.

Check for:

- ✓ Missing values
- ✓ Inconsistent data types (e.g., text in Sales)
- ✓ Outliers (use Column Profile → look at Min/Max)

Click Close & Load to load the cleaned data back to Excel.

3. Basic Summary Statistics in Excel

Add these formulas to new cells:

Metric Formula

Total Sales	=SUM(E2:E9)
Average Sales	=AVERAGE(E2:E9)
Max Sales	=MAX(E2:E9)
Min Sales	=MIN(E2:E9)
Standard Deviation	=STDEV.S(E2:E9)
Count of Transactions	=COUNTA(E2:E9)
Unique Products	=COUNTA(UNIQUE(C2:C9))

4. Create Visuals

Pivot Table

Select your table → Insert → PivotTable.

Drag:

- Region to Rows
- Sales to Values
- Product to Columns (optional)

Charts

Bar Chart: Compare total sales by region.

Histogram: Use a column with numerical values like Sales or Quantity.

Line Chart: Show sales trends over time (Date vs. Sales).

Box Plot: Use a template or manually plot quartiles for sales.

5. Identify Patterns

Look for:

- Which region has the highest/lowest sales?
- Any product consistently underperforming?
- Outliers in Sales or Quantity?
- Correlation between Quantity and Sales?

EDA in Power BI

1. Load and Transform Data

Use Power Query Editor to:

- Clean nulls
- Change data types
- Rename columns
- Create new columns
- Split/merge columns

2. Data Profiling
In Power Query Editor:
Enable Column Quality, Column Distribution, and Column Profile under View tab

These show:

- Count of distinct/empty values
- Value distribution
- Min, max, mean, standard deviation

3. Create Summary Statistics
Use DAX measures or visuals for:

- Mean, Median, Mode
- Variance, Standard Deviation
- Min, Max
- Counts and % distributions

Example DAX:

```
Average Sales = AVERAGE('Sales'[Amount])
```

4. Visualizations
Use these for EDA:

- Bar/Column Charts → category distribution
- Histograms → numeric data distribution (custom binning)
- Box Plots → use custom visuals
- Scatter Plots → correlation analysis
- Matrix/Table Visuals → Pivot-style analysis

5. Filters and Slicers
Use slicers and filters to segment data and observe trends or anomalies in subsets.