Day - 10: Understanding Data and Descriptive Statistics from Theory to Python Implementation

23rd September 2024

Introduction to Data Types:

What is Data?

A collection of Facts, number of qualitative Characteristics that represent information about an event, object or system...

Data is the core building block that allows for the extraction of insights, decision-making, and predictions in Data Science

Why is Data important in Data Science?

- Foundation of Data Science
- Making Data-Driven Decisions
- Transforming Raw Data into Insights

Examples:

- Risk Management
- Fraud Detection
- Inventory Management
- Customer Segmentation
- Renewable Energy Forecasting
- Autonomous Vehicles
- Performance Analysis

Types of Data and their Uses:

Aspect	Data Type	Description	Examples	Representation	Advantage	Disadvantage
Source	Primary Data	Data collected directly by the researchers for specific purpose	Surveys, Experiments, Interviews and sensors	Line Charts, Bar Charts, Histogram	1. Relevance 2. Control over Data Quality	1. Time-Consuming 2. Costly
Source	Secondary Data	Pre-existing data collected for different purpose, repurposed for analysis	Government Reports, Academic Datasets	Heat map, Box Plots, Violin Plot, Area Charts	1. Easily Accessible 2. Cost-Effective	1. Relevance and Fit 2. Quality and Accuracy Concerns 3. Need for Cleaning
Frequency	Frequency Data	Data Showing how often an event or observation occurs...	Customer name, Qualitative Responses	Histogram, Bar Charts, Frequency Distribution	1. Insightful for analysis 2. Facilitates Decision Making	1. Limited Context 2. Potential for Misinterpretation

Aspect	Data Type	Description	Examples	Representation	Advantage	Disadvantage
Frequency	Non-Frequency Data	Data that does not represent counts or occurrences...	Qualitative Nature, Categorization	Contour Plots, Scatter Plots, Violin Plots	1. Rich in Detail 2. Flexibility in Analysis	1. Subjectivity 2. Challenges in Quantification
Nature	Quantitative Data	Numerical information that can be measured and expressed as numbers.	Measurable Values, Types - Discrete and Continuous	Histogram, Scatter Plot, Box plot	1. Statistical Analysis 2. Objective measurement	1. Limited Context 2. Potential Misinterpretation
Nature	Qualitative Data	Consists of Non-Numerical information that represents descriptions, characteristics, or qualities...	Descriptive Nature, Categorization	Bar Chart, Pie Chart, Stacked Bar Chart	1. In-depth insights 2. Flexibility in Approach	1. Subjectivity 2. Challenges in Quantification
Measurement	Discrete Data	Consists of numerical information that can only take specific, distinct values, often representing counts of item values.	Specific Values, Countable Events	Bar Charts, Dot Plots	1. Simplicity in Analysis 2. Clear Categorization	1. Limited Range of Values 2. Potential loss of Information
Measurement	Continuous Data	Refers to numerical information that can take value within a given range.	Range of Values, Measurement Precision	Histograms, Frequency Curves	1. Detailed Analysis 2. Rich Insights	1. Complexity in Interpretation 2. Potential Measurements Errors
Structure	Structured Data	Refers to information thats organized in a defined format like in a matrix (rows and columns)	Sales database, Excel Spreadsheet	Line Chart, Heat Map, Bar Chart	1. Ease of Access and Analysis 2. High Data Integrity	1. Limited Flexibility 2. Inability to Capture Complex Data
Structure	Unstructured Data	Refers to Information that does not follow a specific format or organization	Social Media Posts, Customer Feedback Emails	Word Cloud, Network Graph, Histogram	1. Captures Complex Information Rich Source of Information	1. Complex Analysis 2. Storage and Processing Challenges
Category	Numerical Data	Consists of values represented as numbers. an further be classified in to discrete and continuous types.	Sales Figure, Temperature	Line Charts, Histograms	1. Robust Analysis 2. Detailed Insight	1. Sensitivity to Outliers 2. Complexity
Category	Categorical Data	Represents groups or	Gender, Product	Bar Charts, Pie Charts	1. Simplicity 2. Clear	1. Limited Statistical

Aspect	Data Type	Description	Examples	Representation	Advantage	Disadvantage
		categories of information, often non-numeric in nature. Describes qualitative attributes of characteristics. divided into ordinal and nominal data.	Category		Comparison	Analysis 2. Potential Loss of Information

Common issues of the Data:

Data Type	Common Issues	How to Overcome
Primary Data	1. Time consuming and costly to collect 2. requires well-defined collection to avoid bias 3. May lack historical data for comparison	1. Use Automated Tools for data Collection 2. Implement standardization methods to reduce bias 3. Complement Primary Data with secondary data for historical Events
Secondary Data	1. May not be specific to researcher's needs 2. Could be outdated or irrelevant 3. Often Requires Cleaning and reformatting before use. 4. Potential Issues with data accuracy or trustworthiness	1. Cross-Validate data with other sources 2. Regularly update Datasets 3. Use data cleaning techniques to ensure accuracy 4. Investigate Data origins for reliability
Frequency Data	1. High variability can obscure meaningful patterns 2. Aggregation can lead to data loss 3. difficult to handle with very large dataset	1. Use smoothing techniques like moving averages 2. Keep raw data and provide both detailed and summarized views 3. Implement robust database management systems for large datasets
Non-Frequency Data	1. Harder to apply statistical methods or quantitative analysis 2. Requires categorization for analysis, which can introduce subjectivity	1. Convert non-frequency data into usable formats via classification techniques 2. Apply consistent and transparent categorization criteria to minimize subjectively
Quantitative Data	1. Outliers can skew the results 2. Assumptions like normal distribution may not always hold true. 3. Requires careful handling of missing or incomplete data	1. Use Robust statistics like the median for outlier resistance 2. Test for distribution assumptions and apply appropriate methods 3. Impute or remove missing values based on analysis needs
Qualitative Data	1. Difficult to analyze systematically. 2. Requires thematic coding, which can be subjective 3. Large volumes of data may require advanced tools like Natural Language Processing	1. Use qualitative analysis software for organization 2. Maintain consistent coding practices 3. Utilize AI or NLP tools to process large volumes efficiently
Continuous Data	1. Requires precision in measurement tools 2. Small measurement errors can propagate in calculation 3. Potential difficulty in data aggregation and summarization	1. Use high-quality measurements tools and calibrate regularly 2. Implement error-checking algorithms 3. Use appropriate aggregation techniques that account for precision loss
Discrete Data	1. Can lead to oversimplification of phenomena 2. Limited granularity may miss important leads 3. Not always suitable for predictive models	1. Use hybrid approaches, combining discrete and continuous data when possible 2. Increase Data resolution if needed 3. Choose models better suited for discrete data, like decision tree

Data Type	Common Issues	How to Overcome
Structured Data	1. Inflexible format, requires predefined schema 2. Scaling can be difficult as data grows 3. Cannot easily handle unanticipated variations of data	1. Implement Flexible database structures (NoSQL) 2. Use Modular Data Models 3. Regularly Update schema based on emerging data trends
Unstructured Data	1. Hard to process and analyze without advanced techniques (e.g., text mining, AI) 2. May require extensive storage and processing resources. 3. Quality of insights depends heavily on preprocessing and cleaning.	1. Use AIML Techniques for better processing 2. Invest in scalable cloud storage and processing solutions 3. Implement Rigorous data cleaning protocols before analysis
Categorical Data	1. Risk of oversimplification when grouping complex data 2. May require subjective grouping (e.g., high or low) 3. Harder to quantify relationship between categories	1. Use more nuanced category creation and maintain transparency in groupings 2. Test Category groupings with different thresholds 3. Apply statistical tests for categorical relationships like Chi-square
Numerical Data	1. Assumptions about data distribution might not hold 2. Outliers can heavily influence metrics like the mean 3. Missing values can complicate statistical analysis	1. Use non-parametric tests if assumptions fail 2. Apply outlier detection and removal techniques 3. Impute missing data based on statistical models

Data Visualization:

What is Data Visualization?

It refers to the graphical representation of information and data. It uses elements like charts, graphs, maps, etc. It provides tools to provide an accessible way to understand trends, outliers and patterns in data.

Why is Data Visualization Needed?

- Simplifies Complex Data
- Detect Patterns and Trends
- Communicate Results Effectively
- Quick Decision Making

How does Data Visualization Help us?

- Improves Data Accessibility
- Reveals Hidden Insights
- Data Quality Validation
- Better Understanding or Relationships

Benefits of Data Visualization?

1. Enhanced Comprehension
2. Increased Engagement
3. Decision Support
4. Accessibility for Non- Technical Users
5. Highlighting Key Metrics

Visualization for Exploratory Data Analysis:

EDA or Exploratory Data Analysis is the process of investigating datasets to summarize their main characteristics, often with visual methods.

1. **Initial Data Exploration**
2. **Pattern and Trend Identification**
3. **Understanding Distributions**
4. **Spotting Anomalies**
5. **Feature Selection and Engineering**
6. **Validation of Assumptions**
7. **Clustering and Segmentation**
8. **Comparing Groups or Categories**

Types of Charts and their uses:

Chart Name	Definition	Description	Where to Apply	How to Infer:
Bar Chart	Rectangular bars to represent data values	1. Used to compare categorical data 2. Categories on X axis and values on Y axis 3. Can be plotted vertically or horizontally	Categorical Comparisons (Sales by region, number of product sold)	Taller Bar represents higher values.
Pie Chart	reprsents ata as slices of a circle with each slice representing portion of a whole	1. Used for showing proportions or percentages 2. Each slice size is proportional to the quantity it represents	Representing Parts of a Whole (market share, budget distribution)	The larger the slice, the larger the proportion of that category within the dataset
Histogram	Distribution of a dataset by dividing it into bins and counting the numbers of data points that falls into each pin.	1. Used to understand the distribution of continuous data 2. X-axis represents intervals or bins and Y-axis represents frequency	Examining the distribution of continuous variables like age distribution	The heoight of each bin represents the frequency of values within that range. it helps to assess weathe
Scatter Plot	Uses dots to represent the values of two different variables	1. Visualizes relationships or correlations between two continuous variables 2. Each point represents an observation in the dataset	Exploring relationship between two continuous variables (Height vs Weight, sales vs marketing spend)	A positive correlation shows an upward trend, and a negative correlation shows a downward trend. No visible pattern indicates no correlation
Box Plot	Summarizes data by showing its median, quartiles, and outliers in a graphical format	1. Understand distribution of data, especially its spread, and any outliers 2. Represents the IQR and the whiskers are the outliers showing the extent of the data	Summarizing data distributions like test scores and stock prices	The line inside the box represents the median. Outliers are plotted as individual points. The Whiskers give a sense of the range
Line Plot	Uses line to connect individual data points, typically showing the change of a variable over time	1. Often used to represent time series data 2. X-axis represent time or sequential events and Y-axis represents the values	Tracking changes over time (stock prices, sales trends)	The slope of the line shows whether the variables is increasing, decreasing, or staying constant over time
Violin Plot	Similar to boxplot but rotated kernel density plot on each side.	1. Combines a Box-plot with density estimation 2. Shows the probability density of the data at different values	analyzing distribution and their shapes comparing distributions between categories	The thickness of the plot shows the density of the values, wider part indicates more datapoints

Chart Name	Definition	Description	Where to Apply	How to Infer:
Heat Map	Represents data in matrix form, where values are depicted using color gradients	1. Visualizes correlation Matrices or grid based data 2. Color represents the magnitude of values in the data	Visualizing correlation matrices or distribution of values over 2D space	Darker or more intense colors indicate higher values, lighter colors indicate lower values
Pair Plot	a matrix of scatter plot used t show pairwise elationships between multiple variables in a dataset	each scatter plot shows the relationship between two variables, allowing a quick overview of potential correlations, trends, and distributions	best for EDA to identify trends and correlations between multiple numerical variables	Look for patterns in the scatter plots, such as linear relationships or clustering of points
Stacked Bar Chart	A bar Chart where different categories within each bar are stacked on top of each other	Useful for comparing both the totals and the sub-components(categories) within the data	Use for data with hierarchical categories to compare the total values and the contribution of each sub-category	Compare the height of each bar assess the totals and see how each part contributes to the whole
Boxen Plot	Provides more detailed information about th data distribution especially the tails.	Focuses on large datasets and display more granular information about the distribution , for extremes specifically	Use for large datasets with skewed distributions where you want to visualize data in more details especially outliers and extremes	Similar to box plot but provides deeper insights into the tail
Density Plot	A smooth curve representing the distribution of a continuous variables	Used to estimate the probability density function of a continuous variable, showing where data points are concentrated	Best for visualizing the distribution of continuous data and comparing distribution across different groups	Look for peaks in the curve, which indicate areas of high concentration. Multiple peaks suggest bimodal or multimodal distribution
Area Chart	line chart where area under the line is filled with color to show magnitude	typically used to visualize cumulative frequency over time or between different categories	use when yo want to emphasize the magnitude of the changes over time or across categories	look at the shape of the filled area to understand trends and cumulative totals.
Bubble Chart	A scatter plot where the size of each point represents a third variable	Allows visualization of three variables at once by encoding two variables in the axes and the third in the size of the bubble	Useful when you want to visualize relationships between three variables and show relative importance via bubble size	Larger Bubbles indicate higher values of the third variables, while the position on the axes shows the relationship between the other two variables
Dot Plot	Uses dots to represent individual data points	Useful for visualizing distributions and frequencies in small datasets. Similar to bar chart but more focused on individual points	Ideal for smaller datasets where you want to highlight individual data points and their frequency	The Spread of dots along the axis indicates the distribution and relative frequency of values