Exploratory Data Analysis

What is EDA?

EDA is an approach to analyzing datasets to summarize their main characteristics, often using statistical graphics, plots, and information tables.

It's the first step in your data analysis process, where one formulates theories or hypotheses by visualizing and understanding the patterns, relationships, anomalies, etc., in the data.

Historical Context

- The term "Exploratory Data Analysis" was popularized by John Tukey in the 1970s.

- Tukey emphasized the importance of exploring data before making any assumptions or hypotheses. He argued that data can possess patterns and structures not anticipated, and EDA is essential to discover these.

Why is EDA Important?

01 **Understanding the Nature of the Data:**
Before we can analyze or model the data, we need to know what we're working with. EDA helps us determine the structure, quirks, patterns, and anomalies in the data.

02 **Guiding Model Selection:**
The patterns and relationships discovered during EDA can inform the choice of models or algorithms to apply.

03 **Assumption Validation:**
Many statistical models and machine learning algorithms require assumptions (e.g., linearity, normality). EDA can help verify or challenge these assumptions.

04 **Feature Engineering:**
By understanding the relationships between variables, we can create new features that might improve model performance.

05 **Preventing Costly Mistakes:**
EDA can help detect and handle anomalies and outliers that could lead to incorrect conclusions or poor model performance.

The Iterative Nature of EDA

- EDA is not a linear process; it's iterative. As we explore the data, we may go back and forth, refining our analysis, visualizations, and hypotheses.

- This iterative nature ensures that we continually refine our understanding and approach based on what the data is telling us.

Understanding the Data

Origin of Data:

- Discuss where the data comes from. Is it collected internally, sourced from third parties, or publicly available?
- Mention any potential biases or limitations in the data collection process.

Types of Data

Continuous Data: Numeric data that has an infinite number of possibilities.

Categorical Data: Data that can be divided into distinct categories but doesn't have a numerical
meaning.

Time-Series Data: Data points collected or recorded at specific time intervals.

Text Data: Unstructured data often requiring natural language processing to analyze.

Exploring Data Dimensions

Features VS. Observations

- Features (Columns): Different attributes or measures in the dataset.
- Observations (Rows): Individual data points or samples.

Dataset Size

- Importance of understanding the volume of data. Large datasets might require more computational resources or different analytical approaches, while small datasets might lead to overfitting in modeling.

Data Granularity

- The level of detail in the data. For instance, sales data could be at a transactional level, daily level, or monthly aggregate.

Data Quality: The First Step to Success

Missing Values:

Identify if there are any gaps in the data. Discuss potential reasons and implications.

Duplicates:

Check for and understand the reason behind any duplicate records. Are they true duplicates or repeated measurements?

Data Consistency:

Ensure that data formats and scales are consistent across the dataset. For instance, ensuring all monetary values are in the same currency or all dates are in the same format.

Initial Data Summary:

Using descriptive statistics and basic visualizations to get a feel for the data. This includes measures like mean, median, mode, standard deviation, and simple plots like histograms.

Importance of Documentation (Metadata and Data Dictionary)

- Sometimes datasets come with metadata or a data dictionary that describes each feature in detail. This can include data types, units of measurement, permissible values, and a description of what each feature represents.
- Understanding this can save time and prevent errors in the analysis.

Data Cleaning

Why Data Cleaning Matters:

- Emphasize the adage: "Garbage in, garbage out." Poor quality data can lead to misleading analyses and conclusions.
- Mention that many industry experts claim data cleaning can take up to 80% of the data analysis time.

Framework of Addressing Data Cleaning and Preprocessing

1. Handling Missing Values
2. Removing Duplicates
3. Data Type Conversion
4. Outliers and Anomalies

Handling Missing Values

Identifying Missing Values:

- Use tools or functions to detect NaNs or null values in the dataset.
- Understand the reasons for missing data: Is it random, or is there a pattern?

Strategies to Address Missing Data:

- **Deletion**: Removing rows or columns with missing values. Useful when the proportion of missing data is small.
- **Mean/Median/Mode Imputation:** Filling in with the average, middle, or most frequent value. Suitable for numeric columns.
- **Forward/Backward Fill:** Using the previous or next value to fill gaps, especially useful for time-series data.
- **Model-Based Imputation:** Using models like k-NN or regression to predict and impute missing values.
- **Flagging:** Create a new column to indicate where data was missing.

Treating Duplicates

Identifying Duplicates:

- Using tools or functions to detect repeated rows.
- Understanding how duplicates arose: Were they data entry errors, or are they valid repeated measurements?

Strategies for Handling Duplicates:

- **Deletion:**

  Simply removing the duplicated rows.
- **Aggregation:**

  If duplicates have variations in other columns, consider aggregating data.
- **Verification:**

  In some cases, it might be worth reverting to the data source or collection method to verify whether the duplication is valid.

Data Type Conversion

Ensuring Correct Data Types:

- Data may sometimes be stored in inappropriate formats, e.g., numbers stored as text strings or dates stored as integers.
- Convert data to the most suitable type to facilitate analysis.

Categorical Encoding:

Convert categorical data into a format that can be provided to machine learning algorithms.

- **One-Hot Encoding**
  : Create a new column for each category, with binary indicators.
- **Label Encoding**
  : Assign a unique integer to each category.

Outliers and Anomolies

Detecting Outliers

- Visualization tools like box plots or scatter plots.
- Statistical methods like Z-scores or the IQR method.

Handling Outliers

- **Trimming**
  : Simply removing outlier values.
- **Transformation**
  : Applying mathematical operations to compress extreme values, e.g., log transformation.
- **Imputation**
  : Replacing outliers with statistical measures like mean or median.

Statistical Analysis

Purpose of Statistical Analysis in EDA

- Extract insights from data, summarize main characteristics, and provide a foundation for further analysis or modelling.
- Validate assumptions and hypotheses about the data.

Measure of Central Tendency

Mean

- Sum of all values divided by the number of values.
- Susceptible to extreme values or outliers.

Median

- The middle value when data is sorted in ascending or descending order.
- More robust to outliers compared to the mean.

Mode

- Value that appears most often in the dataset.
- Datasets can be unimodal (one mode), bimodal (two modes), or multimodal (more than two modes).

Measures of Dispersion

Range

- Difference between the max and min values.

Variance

- Average of the squared differences from the mean.
- Provides insight into the data's spread.

Measures of Dispersion

Standard Deviation

- Square root of the variance.
- Measures the average distance between each data point and the mean.

Interquartile Range (IQR)

- Difference between the 75th percentile (Q3) and the 25th percentile (Q1).
- Offers a measure of where the "middle half" of the data lies.

Measures of Shape

Skewness

- Measure of the asymmetry of the probability distribution.
- Positive skew indicates tail on the right, negative skew indicates tail on the left.

Kurtosis

- Measure of the "tailedness" of the probability distribution.
- High kurtosis indicates a sharp peak and fat tails, while low kurtosis indicates a flatter peak and thinner tails.

Correlation Vs. Causation

Correlation

- Quantifies the degree to which two variables change together.
- Correlation coefficient values range from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation.

Causation

- Indicates that a change in one variable is responsible for a change in another.
- Important to emphasize: "Correlation does not imply causation." Just because two variables are correlated does not mean one causes the other.

Visualization Techniques

Purpose of Visualization in EDA

- Transform raw data into a format that can be easily and quickly understood.
- Identify patterns, relationships, anomalies, and more.
- Communicate findings in a manner that's impactful and easy for audiences to grasp.

Univariate Analysis

Continuous Values:

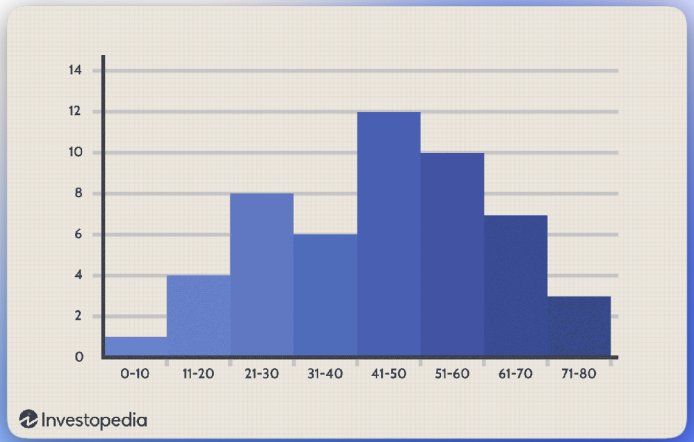
- Histograms
- Box Plots
- Density Plots

Categorical Values:

- Bar Charts
- Pie Charts

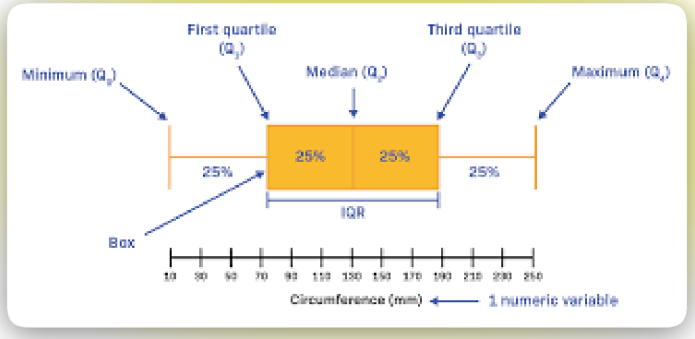
Histograms

- Provides a view of the data's distribution.
- Can identify modes, skewness, and outliers.



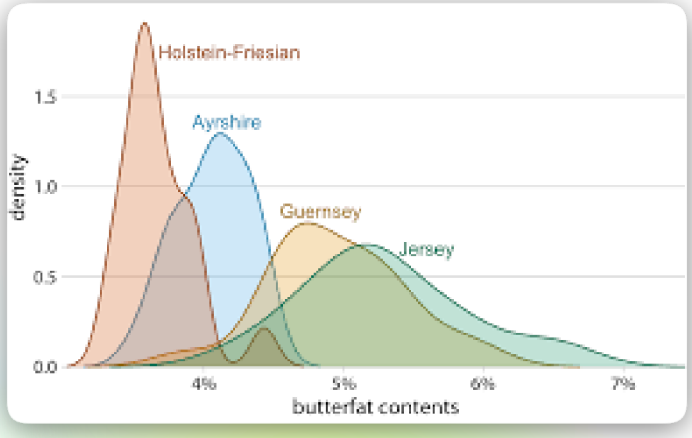
Box Plots

- Displays the distribution's five-number summary: minimum, Q1 (25th percentile), median, Q3 (75th percentile), and maximum.
- Useful for identifying outliers and understanding the data's spread.



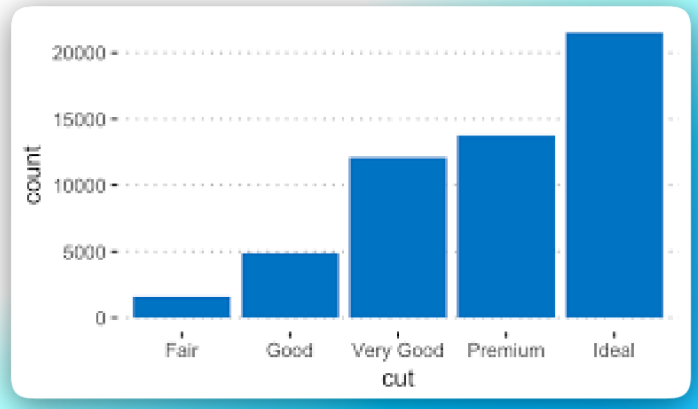
Density Plots

Smoothed version of the histogram, useful for comparing multiple distributions.



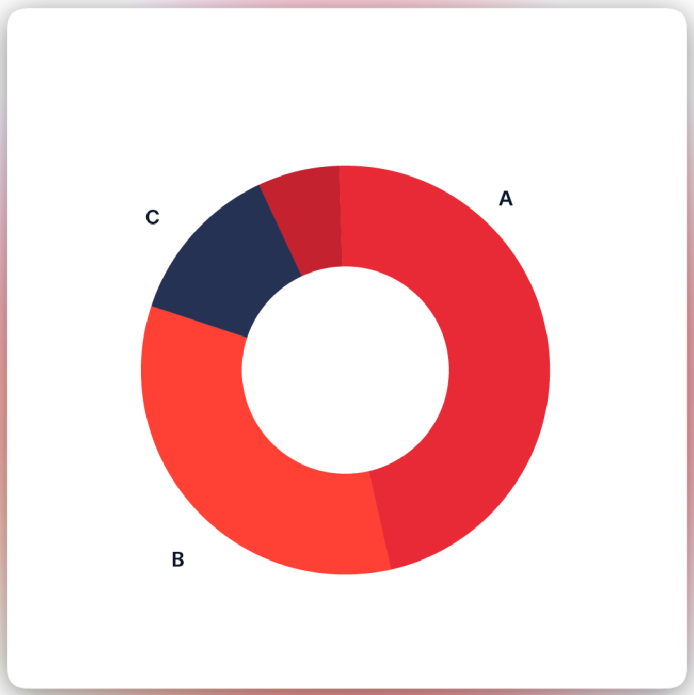
Bar Charts

- Displays the frequency or proportion of categories.
- Can be horizontal or vertical.



Donut Chart

- Represents proportions of categories.
- Best used when there are few categories and the focus is on relative proportions.

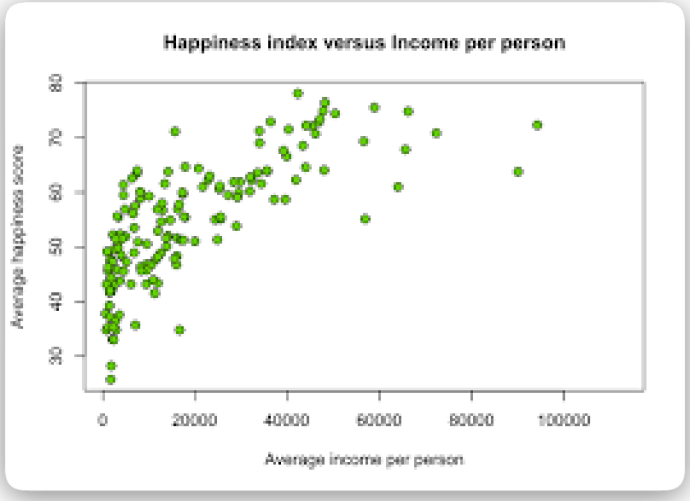


Bivariate Analysis

- Scatter Plots
- Bar Charts
- Box Plots
- Heatmaps

Scatter Plots

- Displays the relationship between two continuous variables.
- Can identify trends, patterns, and correlations.

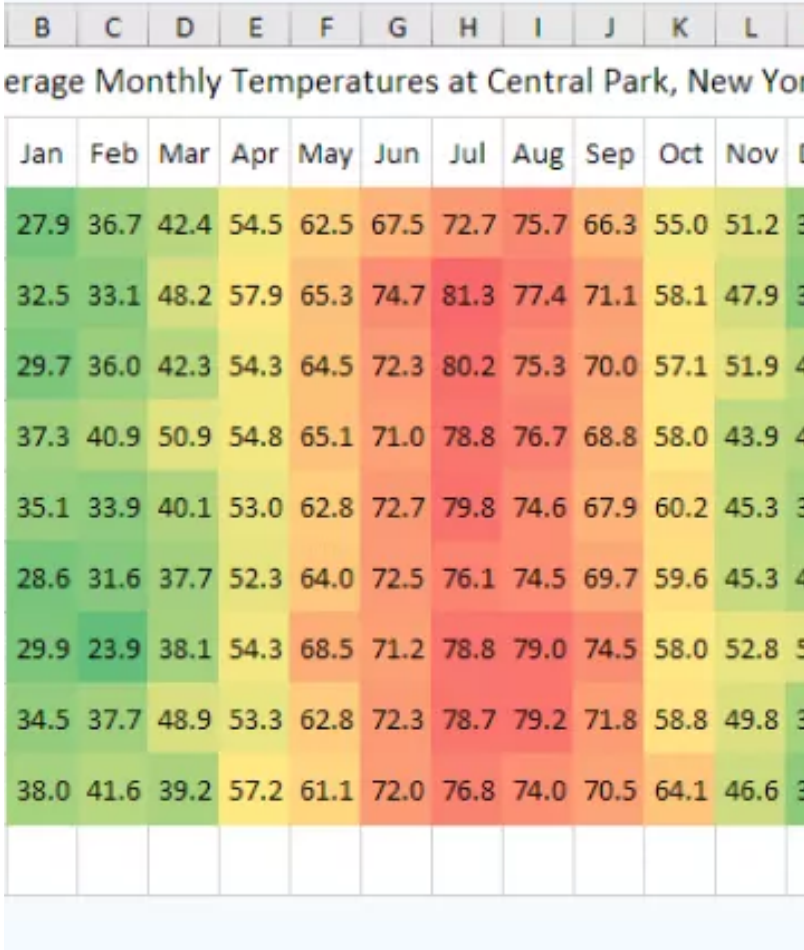


Bar Charts

- For comparing a categorical and a continuous variable or two categorical variables.
- Displays aggregated measures like sums or averages.

Heatmaps

- Useful for representing correlations or data matrices.
- Visualize magnitude using color gradients.



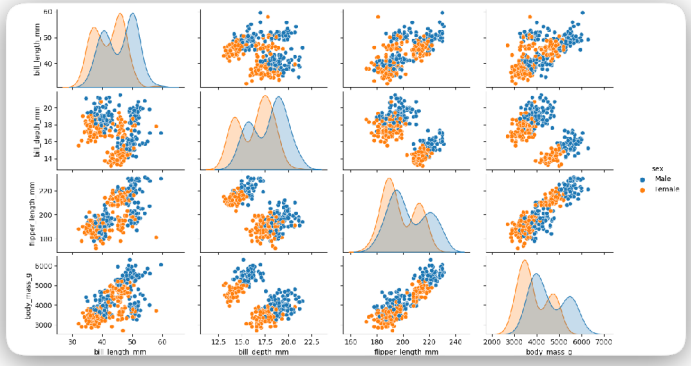
B	C	D	E	F	G	H	I	J	K	L
erage Monthly Temperatures at Central Park, New Yo										
Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov
27.9	36.7	42.4	54.5	62.5	67.5	72.7	75.7	66.3	55.0	51.2
32.5	33.1	48.2	57.9	65.3	74.7	81.3	77.4	71.1	58.1	47.9
29.7	36.0	42.3	54.3	64.5	72.3	80.2	75.3	70.0	57.1	51.9
37.3	40.9	50.9	54.8	65.1	71.0	78.8	76.7	68.8	58.0	43.9
35.1	33.9	40.1	53.0	62.8	72.7	79.8	74.6	67.9	60.2	45.3
28.6	31.6	37.7	52.3	64.0	72.5	76.1	74.5	69.7	59.6	45.3
29.9	23.9	38.1	54.3	68.5	71.2	78.8	79.0	74.5	58.0	52.8
34.5	37.7	48.9	53.3	62.8	72.3	78.7	79.2	71.8	58.8	49.8
38.0	41.6	39.2	57.2	61.1	72.0	76.8	74.0	70.5	64.1	46.6

Multivariate Analysis

- Pair Plots
- Bubble Charts
- Faceted Grids

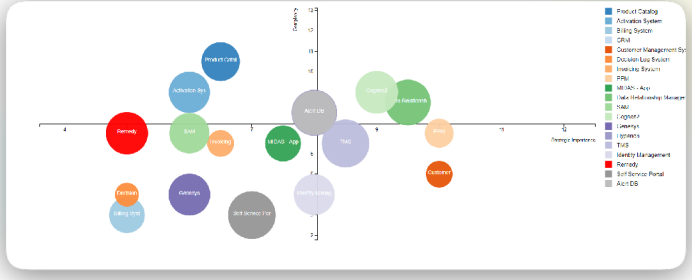
Pair Plots

- Matrix of scatter plots to compare multiple continuous variables.
- Quickly identify relationships and patterns across several variables.



Bubble Charts

- Variation of scatter plots where data points are replaced with bubbles.
- Three variables are represented: x-axis value, y-axis value, and bubble size.



TLDR

01 Understand the Data

02 Clean The Data

03 Look at the statistics of the data

04 Visualize the data