EXPLORATORY DATA ANALYSIS

Intro to Machine Learning

WHAT IS EDA?

- The analysis of datasets based on various numerical methods and graphical tools.
- Exploring data for patterns, trends, underlying structure, deviations from the trend, anomalies and strange structures.
- It facilitates discovering unexpected as well as conforming the expected.
- Another definition: An approach/philosophy for data analysis that employs a variety of techniques (mostly graphical).

Objective of EDA

- Maximize insight into a dataset
- Uncover underlying structure
- Extract important variables
- Detect outliers and anomalies
- Test underlying assumptions
- Develop valid models
- Determine optimal factor settings (Xs)

Objective of EDA

The goal of EDA is to open-mindedly explore data.

Tukey: *EDA is detective work… Unless detective finds the clues, judge or jury has nothing to consider.*

Steps of EDA

- Generate good research questions
- Data restructuring: You may need to make new variables from the existing ones.
  - Instead of using two variables, obtaining rates or percentages of them
  - Creating dummy variables for categorical variables
- Based on the research questions, use appropriate graphical tools and obtain descriptive statistics. Try to understand the data structure, relationships, anomalies, unexpected behaviors.
- Try to identify confounding variables, interaction relations and multicollinearity, if any.
- Handle missing observations
- Decide on the need of transformation (on response and/or explanatory variables).
- Decide on the hypothesis based on your research questions

Classification of EDA

- Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate).
- Non-graphical methods generally involve calculation of summary statistics, while graphical methods obviously summarize the data in a diagrammatic or pictorial way.
- Univariate methods look at one variable (data column) at a time, while multivariate methods look at two or more variables at a time to explore relationships. Usually our multivariate EDA will be bivariate (looking at exactly two variables), but occasionally it will involve three or more variables.
- It is almost always a good idea to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.

Example

9 variables fro 329 metropolitan areas

- Climate mildness
- Housing cost
- Health care and environment
- Crime
- Transportation supply
- Educational opportunities and effort
- Arts and culture facilities
- Recreational opportunities
- Personal economic outlook

+ latitude and longitude of each city

Questions:

- How is climate related to location?
- Are there clusters in the data (excluding location)?
- Are nearby cities similar?
- Any relation bw economic outlook and crime?
- What else???

Data Types and Measurement Scales

Variables may be one of several types, and have a defined set of valid values.

Two main classes of variables are:

- **Continuous Variables**: (Quantitative, numeric).
  - Continuous data can be rounded or \binned to create categorical data.
- **Categorical Variables**: (Discrete, qualitative).
  - Some categorical variables (e.g. counts) are sometimes treated as continuous.

Categorical Data

- Unordered categorical data (nominal)
    - 2 possible values (binary or dichotomous)
        - Examples: male/female, alive/dead, yes/no.
    - Greater than 2 possible values - No order to categories
        - Examples: colors, religion, nationality.

- Ordered categorical data (ordinal)
    - Ratings or preferences
    - Education level: high school < bachelor’s < master’s < PhD
    - Customer satisfaction: very dissatisfied < dissatisfied < neutral < satisfied < very satisfied
    - T-shirt sizes: small < medium < large < extra-large

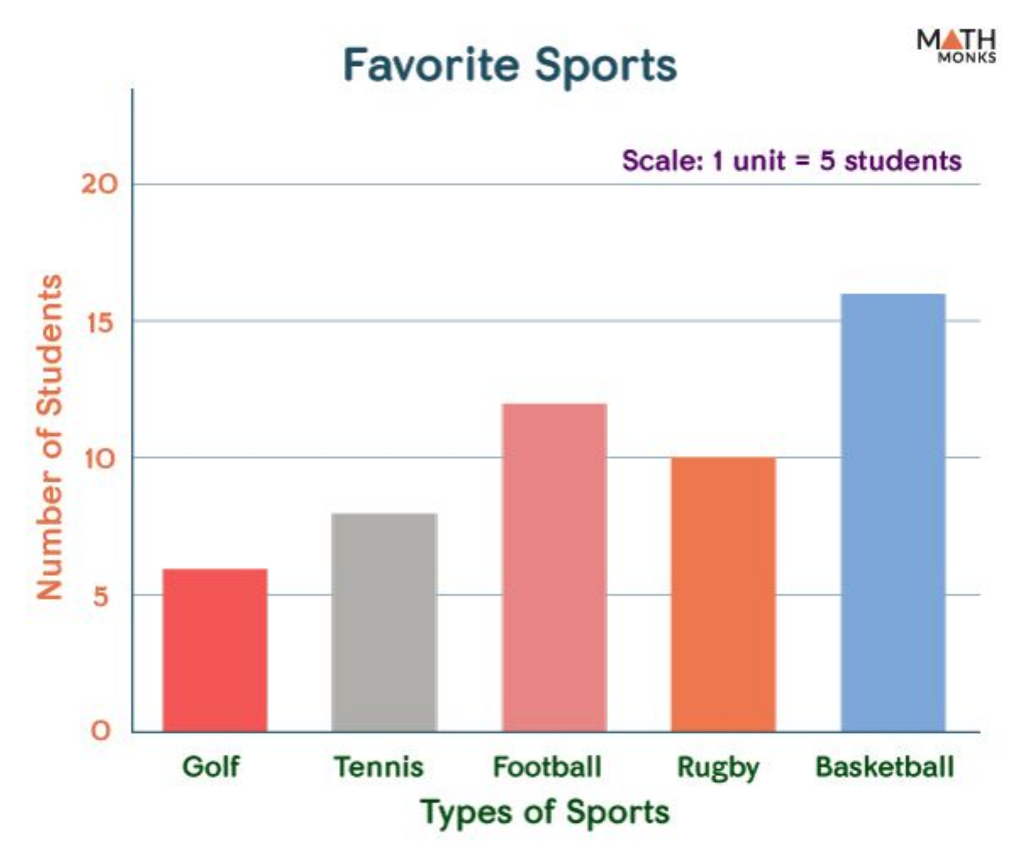
Summarizing Data With Tables and Plots

Bar Chart

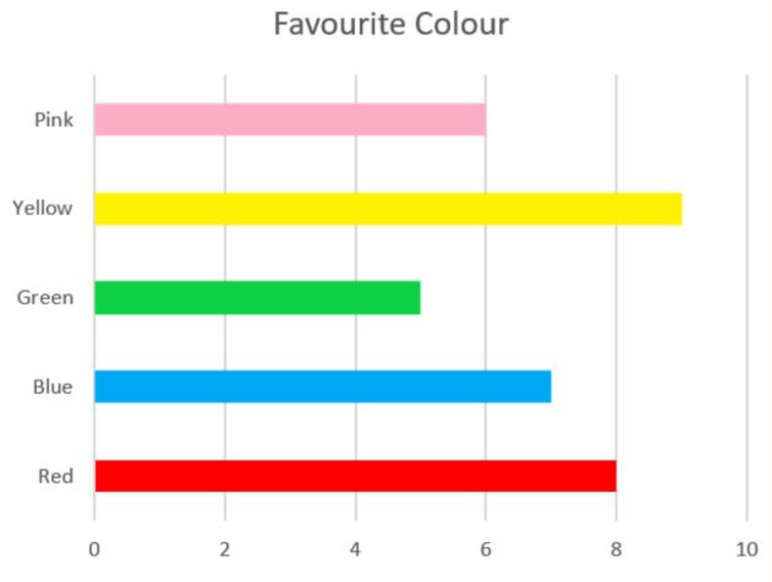
- Data Type: Categorical

- Use: Displays counts or values across distinct categories.

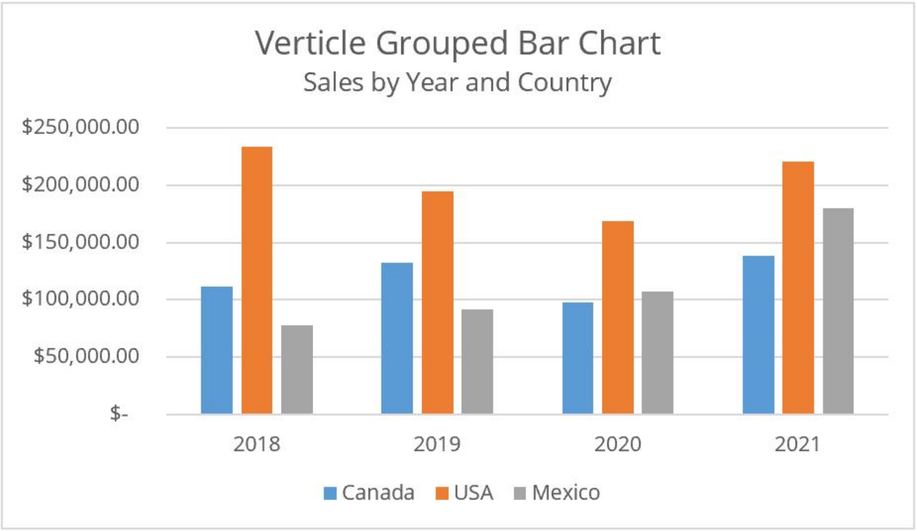
➢ Easy way to compare category performance or frequency.



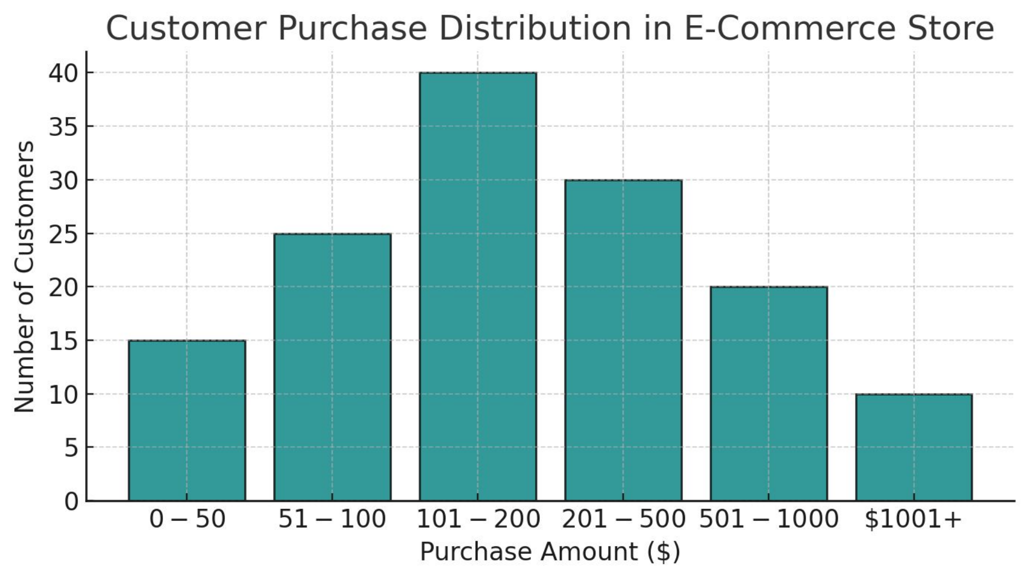
Bar Chart



Favourite Colour
Pink
Yellow
Green
Blue
Red
0
2
4
6
8
10



Verticle Grouped Bar Chart
Sales by Year and Country
$250,000.00
$200,000.00
$150,000.00
$100,000.00
$50,000.00
$-
2018
2019
2020
2021
Canada
USA
Mexico



Histogram

- Data Type: Continuous
- Use: Visualizes the distribution of a numeric variable by grouping data into bins.

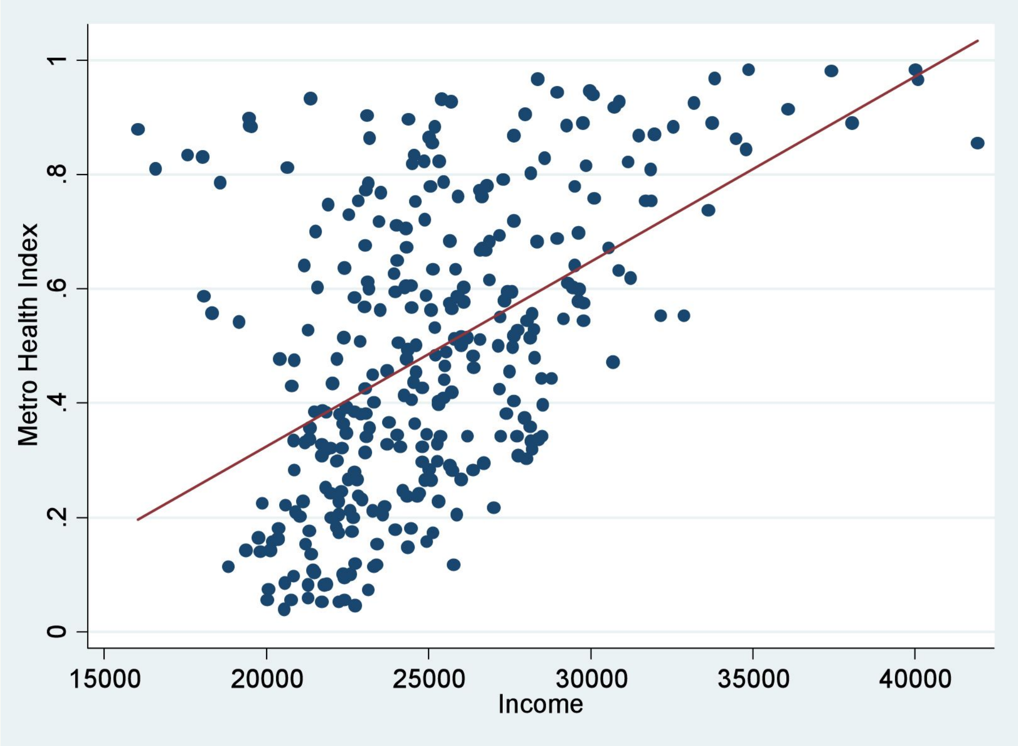
➤ Helps detect skewness, peaks, and spread.

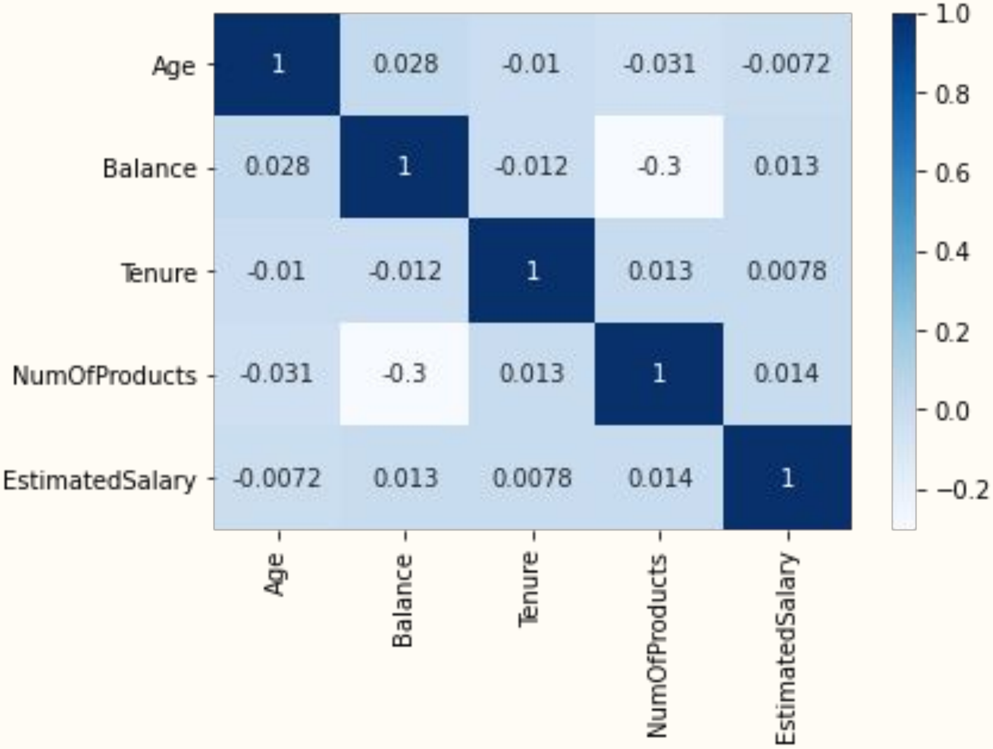
Scatter Plot

- Data Type: Continuous

- Use: Plots the relationship or correlation between two numeric variables.

➤ Reveals trends, clusters, or outliers.





	Age	Balance	Tenure	NumOfProducts	EstimatedSalary
Age	1	0.028	-0.01	-0.031	-0.0072
Balance	0.028	1	-0.012	-0.3	0.013
Tenure	-0.01	-0.012	1	0.013	0.0078
NumOfProducts	-0.031	-0.3	0.013	1	0.014
EstimatedSalary	-0.0072	0.013	0.0078	0.014	1

Correlation Heatmap

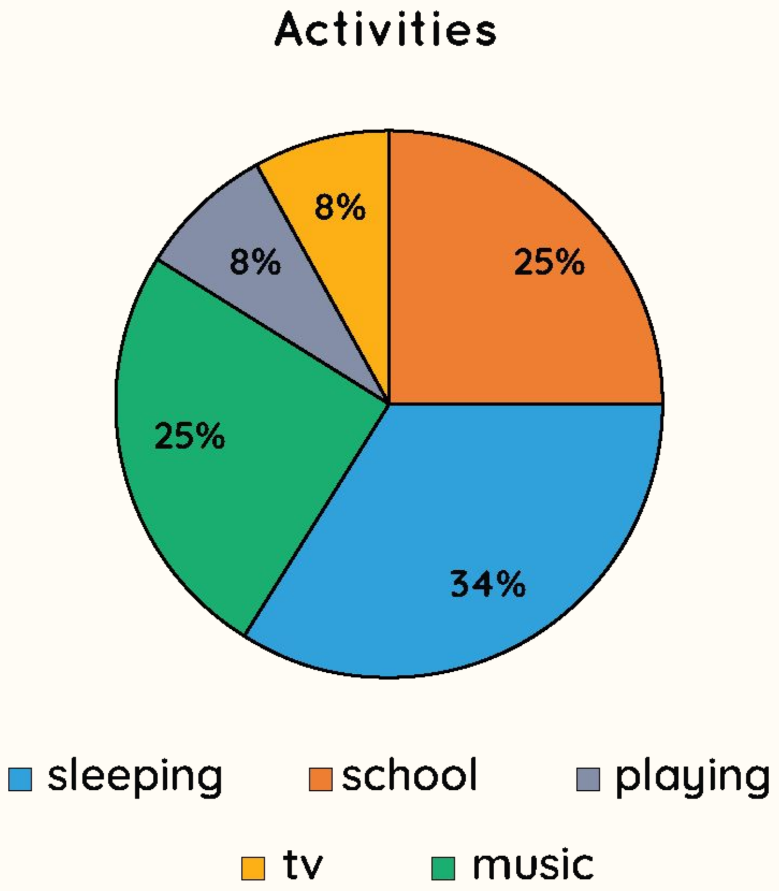
- Data Type: Continuous (matrix)
- Use: Color-coded matrix showing correlation coefficients between features.

➤ Helps detect multicollinearity or strong linear relationships.

Pie Chart

- Data Type: Categorical
- Use: Circular chart divided into slices to illustrate numerical proportions of categories.

➢ Helps visualize the relative percentage or share of categories within a whole. Useful when comparing a small number of categories.



Thank You