Unit – I

Introduction to EDA

Syllabus
EDA: - Definition, need, steps. Introduction to Dataset: - Definition, Variables and their types, Identify numerical and categorical variables, Cardinality in categorical variables, Relationship between variables, Covariance and Correlation, concept of multicollinearity

1.1. Definition

Exploratory Data Analysis (EDA) is the process of analyzing and summarizing a dataset to understand its main characteristics, uncover patterns, detect anomalies, and test hypotheses—primarily through visual methods and simple statistics—before applying formal modeling or machine learning techniques.

1.2. Need for EDA

a. **Understanding the Data:**
Before building any model or drawing conclusions, it's crucial to get a clear picture of the dataset—its size, types of variables, distributions, and quality.

b. **Detecting Errors and Missing Values:**
Real-world data is often messy. EDA helps identify errors, missing values, and inconsistencies that need to be cleaned or addressed.

c. **Identifying Patterns and Relationships:**
Helps uncover trends, correlations, and associations between variables which can guide further analysis or feature selection.

d. **Spotting Outliers and Anomalies:**
Outliers can skew results or indicate interesting phenomena. EDA detects these unusual data points.

e. **Validating Assumptions:**
Many statistical models assume certain data properties (like normal distribution). EDA checks if these assumptions hold.

f. **Guiding Model Selection and Feature Engineering:**
Insights gained from EDA help decide which algorithms might work best and which features should be created, transformed, or removed.

g. **Improving Communication:**
Visualizations and summaries from EDA make it easier to explain the data story to stakeholders.

1.3. Steps in Exploratory Data Analysis (EDA)

Exploratory Data Analysis involves a systematic process to understand and prepare the data before modeling. It typically includes collecting and cleaning the data, exploring patterns through visualization and statistics, and transforming features to improve analysis outcomes.

1.3.1. Data Collection and Import

- Collect data from various sources like CSV files, databases, APIs, or web scraping. Then import it into your working environment.
- **Tools:**
  - Python: pandas.read_csv(), SQLAlchemy for databases

  - R: read.csv(), DBI package
  - Excel or Google Sheets
- **Example:** Load a sales dataset from a CSV file into a pandas DataFrame in Python.

1.3.2. Data Cleaning and Preprocessing

- Address missing data (imputation or removal), fix inconsistencies (e.g., date formats), remove duplicates, and filter out irrelevant data.
- **Tools & Techniques:**
  - Python: pandas.isnull(), dropna(), fillna(), duplicated()
  - R: na.omit(), tidyr::fill(), dplyr::filter()
- **Example:** Replace missing values in the "Age" column with the median age or remove rows with too many missing fields.

1.3.3. Data Exploration and Visualization

- Summarize the data using statistics and visualize distributions and relationships to spot patterns and outliers.
- **Common Techniques:**
  - Summary statistics: mean, median, standard deviation
  - Visualizations: Histograms, box plots, scatter plots, heatmaps
- **Tools:**
  - Python: matplotlib, seaborn, pandas.describe()
  - R: ggplot2, summary()
  - Tableau, Power BI for interactive visuals
- **Example:** Plot a histogram of customer ages to see the distribution or a scatter plot to check correlation between advertising spend and sales.

1.3.4. Feature Engineering and Transformation

- Create new variables (features), transform existing ones (e.g., normalize or encode categorical data), and select features to improve model accuracy.
- **Techniques & Tools:**
  - Encoding: One-hot encoding, label encoding (pandas.get_dummies(), sklearn.preprocessing)
  - Scaling: StandardScaler, MinMaxScaler (from sklearn)
  - Feature selection: Correlation analysis, Recursive Feature Elimination (RFE)
- **Example:** Convert a "Date of Purchase" column into "Day of Week" or "Month" to capture seasonal trends.

1.4. Variables and their types

In data analysis, **variables** are the characteristics or properties that can take on different values across observations or subjects. They represent the data attributes that you measure, observe, or record about an individual, event, or object.

1.4.1. Types of Variables

1. **Numerical Variables (Quantitative):**
   These variables express quantities and are represented by numbers. They allow mathematical operations like addition, subtraction, averaging, etc.
   - **Continuous variables:** Can take any value within a range and often represent measurements. For example, height can be 170.5 cm, 171 cm, or any value in between.
   - **Discrete variables:** Can only take specific, separate values, usually counts. For example, the number of children a person has can only be whole numbers (0, 1, 2, etc.).
2. **Categorical Variables (Qualitative):**
   These variables represent categories or groups rather than numbers. They describe qualities or labels.
   - **Nominal variables:** Categories without any inherent order. For example, types of fruits (apple, banana, cherry) or gender (male, female, non-binary). You can count or group by these but cannot rank them.
   - **Ordinal variables:** Categories with a meaningful order or ranking but the intervals between ranks are not necessarily equal. For example, customer satisfaction levels (satisfied, neutral, dissatisfied) or education level (high school < bachelor < master). These have a natural sequence but the difference between categories isn't uniform.

1.4.2. How to Identify Numerical vs. Categorical Variables

- **Numerical Variables:**
  - Contain numeric values.
  - You can perform arithmetic calculations on them (mean, sum, etc.).
  - Examples: age, temperature, income, test scores.
- **Categorical Variables:**
  - Contain labels or names.
  - Arithmetic calculations don't apply; instead, you count frequencies or group them.
  - Examples: blood type, city name, product category.

Variable Name	Type	Example Values	Notes
Age	Numerical (Discrete)	23, 45, 30	Counts years, whole numbers only
Height	Numerical (Continuous)	170.5 cm, 180.2 cm	Measurements can have decimals
Gender	Categorical (Nominal)	Male, Female, Other	No order, just groups
Education Level	Categorical (Ordinal)	High School < Bachelor < Master	Ordered categories
Number of Pets	Numerical (Discrete)	0, 1, 2	Count of pets

1.5. Cardinality in categorical variables

Cardinality in the context of categorical variables refers to the number of unique categories or distinct values that a categorical variable can take. It essentially measures the “variety” or “diversity” within that variable.

1.5.1. Types of Cardinality

1. **Low Cardinality:**
   Variables with only a small number of unique categories.
   - Easy to encode and analyze.
   - Example: Gender (Male, Female), Yes/No responses.
2. **Medium Cardinality:**
   Variables with a moderate number of unique categories.
   - May require some special handling like grouping or frequency encoding.
   - Example: Country names (~200 countries worldwide).
3. **High Cardinality:**
   Variables with many unique categories, often in hundreds, thousands, or more.
   - Can cause problems in encoding (curse of dimensionality).
   - Might lead to sparse data and overfitting if not handled properly.
   - Example: Customer IDs, product SKUs, ZIP codes.

Example:

CustomerID	Gender	Country	FavoriteProduct	PurchaseAmount
1001	Male	USA	Laptop	1200
1002	Female	Canada	Smartphone	800
1003	Male	USA	Headphones	150
1004	Female	UK	Laptop	1300
1005	Male	India	Smartphone	700
1006	Female	USA	Tablet	600
1007	Male	Germany	Laptop	1100
...	...	...	...	...

- ✓ **Gender:**
  - Categories: Male, Female
  - Cardinality: 2 (Low)
- ✓ **Country:**
  - Categories: USA, Canada, UK, India, Germany, ...
  - Cardinality: ~200 (Medium)
- ✓ **FavoriteProduct:**
  - Categories: Laptop, Smartphone, Headphones, Tablet, Camera, … potentially dozens or hundreds.
  - Cardinality: Medium to High depending on product range.
- ✓ **CustomerID:**
  - Unique for every customer, potentially thousands or millions.
  - Cardinality: Very High

1.5.2. Impact of cardinality

Impact of Cardinality on Data Analysis and Modeling

1. Model Complexity and Dimensionality

- **High cardinality** categorical variables can significantly increase the number of features after encoding, especially when using one-hot encoding.
- For example, a variable with 1,000 unique categories turns into 1,000 new binary columns, which increases the dimensionality of the dataset drastically.
- This leads to the **curse of dimensionality**, making the model more complex and harder to train.

2. Computational Efficiency

- More features mean more computation. Models take longer to train and require more memory.
- High cardinality variables can slow down the entire machine learning pipeline, from preprocessing to training and prediction.

3. Risk of Overfitting

- High cardinality often leads to sparse data — many categories have only a few samples.
- Models may overfit on these rare categories because the model “memorizes” noise or peculiarities of those few examples, reducing generalization performance on unseen data.

4. Challenges in Interpretation

- Models built on high-cardinality features encoded with one-hot or similar methods become harder to interpret because there are so many dummy variables.
- It becomes challenging to understand the importance or effect of each category.

5. Visualization Difficulties

- Visualizing high-cardinality categorical variables is tough. Bar plots or frequency charts become cluttered and unreadable when categories are too many.
- This makes exploratory data analysis (EDA) less effective.

6. Data Quality Issues

- Variables with high cardinality might contain noisy, inconsistent, or irrelevant categories (e.g., misspellings or rare categories).
- This noise can degrade model performance if not handled properly.

1.5.3. How to Handle High Cardinality Variables

- ✓ **Grouping Rare Categories:** Combine less frequent categories into a single “Other” group to reduce cardinality.
- ✓ **Frequency Encoding:** Replace categories with their frequency counts.
- ✓ **Target Encoding:** Replace categories with the average target variable value for that category (useful in supervised learning).
- ✓ **Embedding:** Use learned dense vector representations (common in deep learning).
- ✓ **Hashing Trick:** Hash categories into fixed-size buckets to limit dimensionality.

1.6. Relationship between variables

In data analysis, understanding the relationship between variables is key to uncovering insights, building predictive models, and making decisions. Relationships describe how one variable change when another variable changes.

1.6.1. Types of Relationships Between Variables

1. **No Relationship (Independence):**
   Changes in one variable do not affect the other. They behave independently.

2. **Positive Relationship (Positive Correlation):**
   As one variable increases, the other tends to increase as well.
   - Example: Height and weight usually have a positive relationship.
3. **Negative Relationship (Negative Correlation):**
   As one variable increases, the other tends to decrease.
   - Example: The number of hours spent watching TV might negatively relate to test scores.
4. **Non-linear Relationship:**
   Variables relate in a pattern that isn't a straight line (e.g., quadratic, exponential).
   - Example: Stress and performance might have an inverted U-shaped relationship.

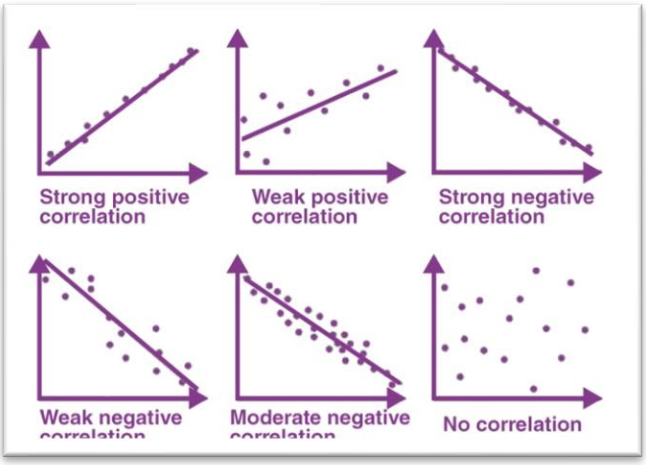


Figure 1. Relationship between variables

1.6.2. How to Measure Relationships

✓ **For Numerical Variables:**
- **Correlation Coefficient (Pearson's r):** Measures strength and direction of linear relationship (-1 to 1).
- **Spearman's Rank Correlation:** For monotonic relationships, even if non-linear.

✓ **For Categorical Variables:**
- **Chi-Square Test:** Tests if two categorical variables are independent.
- **Cramér's V:** Measures association strength between categorical variables.

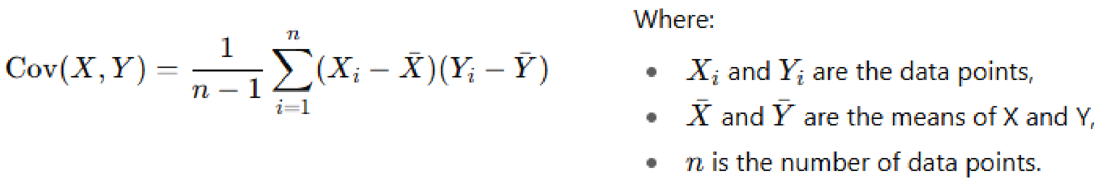
✓ **Between Numerical and Categorical Variables:**
- **Box plots:** Visualize distribution of numerical variable for each category.
- **ANOVA (Analysis of Variance):** Tests if means differ significantly across categories.

1.7. Covariance and Correlation

1.7.1. Covariance

- **Definition:**
  Covariance measures the direction of the linear relationship between two numerical variables. It tells you whether variables tend to increase or decrease together.
- **Interpretation:**
  - **Positive covariance:** When one variable increases, the other tends to increase.
  - **Negative covariance:** When one variable increases, the other tends to decrease.

  - **Covariance close to zero:** Little or no linear relationship.
- **Formula:**



$$\mathrm{Cov}(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$$

Where:

- X_i and Y_i are the data points,
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y,
- n is the number of data points.

Interpretation:

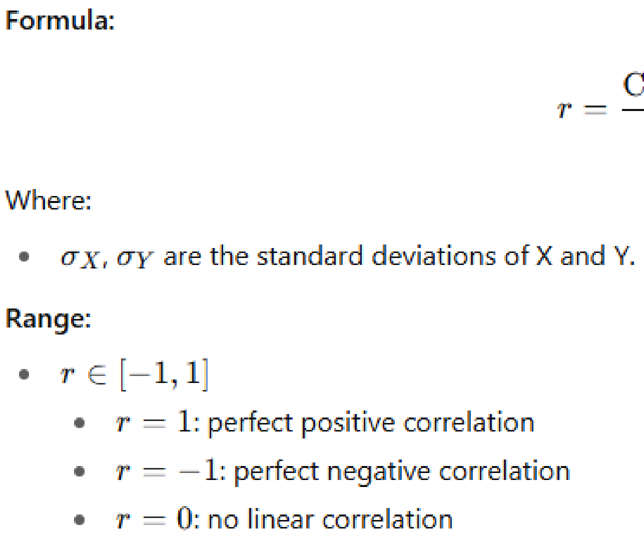
- Covariance is **scale-dependent**, so its magnitude is not easily interpretable.
- Units are the product of the units of X and Y.

1.7.2. Correlation

Definition:

Correlation measures both the **strength and direction** of the linear relationship between two variables.

- The most common is the **Pearson correlation coefficient (r)**.
- It's a **normalized version of covariance**



Formula:

$$r=\frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

Where:

- σ_X, σ_Y are the standard deviations of X and Y.

Range:

- $r \in [-1,1]$
  - $r=1$: perfect positive correlation
  - $r=-1$: perfect negative correlation
  - $r=0$: no linear correlation

Interpretation:

- Correlation is **scale-invariant** — it doesn't depend on the units.
- It's easier to interpret than covariance.

1.8. Concept of multicollinearity

- ✓ Multicollinearity refers to a situation in regression models where two or more independent variables (predictors) are highly correlated with each other.
- ✓ In other words, they contain redundant information, which can make it difficult to interpret the effect of each individual variable on the dependent variable.
- ✓ This redundancy can lead to unstable estimates of the regression coefficients.

1.8.1. Why is Multicollinearity a Problem?

1. **Unstable Coefficients**:
   When predictors are highly correlated, the regression model may struggle to determine how much of the variability in the dependent variable is due to each individual predictor. This can result in very large or very small coefficient estimates that don't reflect the true relationship between the predictors and the target variable.
2. **Inflated Standard Errors**:
   Multicollinearity can inflate the standard errors of the coefficients, which means you might fail to reject the null hypothesis (i.e., your predictors might look insignificant even if they're important).
3. **Overfitting**:
   A model with multicollinearity can become overly sensitive to small changes in the data, leading to overfitting where the model fits the training data very well but performs poorly on new data.

1.8.2. Detecting Multicollinearity

a. **Correlation Matrix**:
The simplest way to detect multicollinearity is by checking the **correlation matrix** of the predictors. High correlations between predictors (typically above 0.8 or 0.9) suggest multicollinearity.

b. **Variance Inflation Factor (VIF)**:
The **VIF** quantifies how much the variance of the estimated regression coefficients is inflated due to multicollinearity. It's calculated for each predictor in the model:

c. **Condition Number**:
This measures the stability of the matrix used to calculate the regression coefficients. High condition numbers (typically above 30) suggest multicollinearity.

1.8.3. Effects of Multicollinearity:

Without Multicollinearity:

Predictor	Coefficient Estimate	Std Error	t-Statistic
X1	0.5	0.1	5.0
X2	2.0	0.4	5.0

With Multicollinearity:

Predictor	Coefficient Estimate	Std Error	t-Statistic
X1	0.2	0.8	0.25
X2	1.5	1.5	1.0

In the second case, the coefficients are **unstable** (larger standard errors), making it harder to determine the actual effect of each predictor.

****

Review Questions:

1. Define Exploratory Data Analysis (EDA). State any four needs of performing EDA.
2. Explain the steps involved in Exploratory Data Analysis (EDA) with examples.
3. Classify the following variables into Numerical (Continuous/Discrete) and Categorical (Nominal/Ordinal): Height, Education level, Age in years, Blood group.
4. Differentiate between covariance and correlation with respect to interpretation and scale.
5. Explain Cardinality in categorical variables. Give examples of low, medium, and high cardinality.
6. Define Exploratory Data Analytics (EDA) and explain any three needs for EDA.
7. A company collects customer data including Gender, Country, Customer ID, and Favorite Product. Identify which variables are low, medium, or high cardinality.
8. Discuss the impact of high cardinality variables on data analysis and modeling.
9. Evaluate different techniques to handle high cardinality variables in datasets.
10. Explain steps involved in EDA along with tools & example.
11. How can you measure relationships between variables? Explain with examples for numerical vs numerical and categorical vs numerical cases.
12. Describe numerical and categorical variables in detail.
13. Explain covariance and correlation in terms of definition, formulae and interpretation.
14. What is multicollinearity? Why is it a problem in regression models? How can it be detected?
15. Describe different types of variables in data analysis with suitable examples.
16. Explain the difference between numerical variables (continuous vs discrete) and categorical variables (nominal vs ordinal) with examples.
17. What is cardinality? Explain types of cardinality with an example.
18. Illustrate with examples how outliers and anomalies can be identified during Exploratory Data Analysis.
19. Explain the impact of cardinality on data analysis and data modelling.
20. Compare the role of covariance, correlation, and multicollinearity in understanding relationships between variables.
21. Explain different types of relationship between variables and also describe how to measure relationship between the variables.
22. Explain why multicollinearity is a problem and also explain how to detect multicollinearity.
23. Explain how data cleaning and preprocessing steps improve the quality of Exploratory Data Analysis.

****