Day 11 – Statistical Analysis & ML Prep

What is Descriptive Statistics?

- *Summarizes and describes data*
- *Helps understand data distribution*
- *First step before analysis or ML*
- *Answers: What does the data look like?*

Key Measures

- Central Tendency

•Mean

•Median

•Mode

- Dispersion

•Variance

•Standard Deviation

•Range

Why It Matters

- *Detect outliers*
- *Compare groups (weekday vs weekend)*
- *Validate data quality*
- *Foundation for hypothesis testing & ML*

Hypothesis Testing

- *Statistical method to test assumptions*
- *Compare two groups or conditions*
- *Uses sample data to infer population behavior*

Key Components

- *Hypothesis (H_0) → No difference*
- *Alternative Hypothesis (H_1) → Significant difference*
- *p-value → Probability of results under H_0*
- *Significance level (α) → Usually 0.05*

Example (Weekday vs Weekend)

- H_0: Average sales are same on weekdays & weekends
- H_1: Average sales differ
- Perform t-test
- Decision based on p-value

A/B Test Design

- *Experiment comparing two variants (A & B)*
- *Measures impact of a change*
- *Widely used in product & marketing analytics*

A/B Test Structure

- *Control group (A)*
- *Treatment group (B)*
- *Single variable change*
- *Random user assignment*

Metrics & Evaluation

- *Conversion rate*
- *Revenue per user*
- *Engagement metrics*
- *Statistical significance check*

What is Feature Engineering?

- *Transform raw data into useful features*
- *Improves ML model performance*
- *Combines domain knowledge + data check*

Common Techniques

- *Date features (day, month, weekend)*
- *Aggregations (avg sales per user)*
- *Encoding categorical variables*
- *Scaling & normalization*

Why Feature Engineering is Critical

- *Better patterns for ML models*
- *Reduces noise*
- *Increases accuracy*
- *Often more important than model choice*