**Types of T-Tests**:

**Steps in Hypothesis Testing**:

1. State H0 and H1 (e.g., H0: μ = 50, H1: μ ≠ 50).
2. Choose α (e.g., 0.05).
3. Calculate test statistic (t = (M - μ0)/(s/√n) for one-sample).
4. Find p-value using t-distribution.
5. Compare p to α:
   * p < α: Reject H0, evidence for H1.
   * p ≥ α: Fail to reject H0, insufficient evidence.

**Interpreting T-Test Results**:

* Small p-value (< 0.05): Statistically significant, reject H0.
* Large |t-statistic|: Larger difference between groups; influenced by sample size.
* Example: t = 2.8, p = 0.01, df = 20 → reject H0; check effect size (e.g., d = 0.6 → medium effect).
* Practical significance: A small p doesn’t always mean a meaningful difference (e.g., tiny difference in large samples).

**Assumptions for T-Tests**:

* **Normality**: Data approximately normal (check via Shapiro-Wilk test or Q-Q plot); robust for n > 30.
* **Homogeneity of variances** (independent samples): Variances equal (check via Levene’s test).
* **Independence**: Observations independent (e.g., no paired data in independent test).
* Violation fixes: Use Welch’s t-test (unequal variances) or non-parametric tests (non-normal).

**One-Tailed vs. Two-Tailed Tests**:

* **One-tailed**: Directional H1 (e.g., H1: μ > 50, “new method improves scores”).
* **Two-tailed**: Non-directional H1 (e.g., H1: μ ≠ 50, “method changes scores”).
* Example: Testing if a drug increases heart rate → one-tailed; testing if it affects heart rate → two-tailed.

**Hypothesis for Correlation**:

* H0: r = 0 (no correlation).
* H1: r ≠ 0 (correlation exists).
* Test: t = r√(n-2) / √(1-r²); compare to t-distribution.

**Correlation ≠ Causation**:

* Third variables, reverse causation, or spurious relationships may explain correlation.
* Example: Ice cream sales and drowning rates correlate (r = 0.8) due to summer season, not causation.
* **One-sample**: Compare sample mean to a known/population value (e.g., is average test score = 70?).
* **Independent samples**: Compare means of two independent groups (e.g., male vs. female salaries).
* **Paired samples**: Compare means from the same group at different times (e.g., pre- vs. post-training scores).

**Key Definitions**:

* **Null Hypothesis (H0)**: Default assumption, no effect/difference (e.g., μ = μ0).
* **Alternative Hypothesis (H1)**: Competing claim, there is an effect/difference (e.g., μ ≠ μ0).
* **P-value**: Probability of observing results (or more extreme) if H0 is true; small p (< α) suggests rejecting H0.
* **T-statistic**: Standardized difference between sample and population mean; |t| large = stronger evidence against H0.
* **Degrees of Freedom (df)**: Values free to vary; one-sample: n-1; independent: n1+n2-2; paired: n-1.
* **Significance Level (α)**: Threshold to reject H0 (typically 0.05 = 5% chance of Type I error).
* **Effect Size**: Practical significance (e.g., Cohen’s d = (M1-M2)/SD; small = 0.2, medium = 0.5, large = 0.8).

**Multicollinearity**:

* High correlation (|r| > 0.7) between predictors in regression; inflates variance, unstable estimates.
* Example: citympg vs. highwaympg, r = 0.9 → multicollinearity risk; consider dropping one variable.
* Check: Correlation matrix or Variance Inflation Factor (VIF > 5 indicates issue).

**Correlation Concepts**:

* **Correlation**: Measures strength/direction of linear relationship between two variables.
* **Pearson’s r**: Interval/ratio data, assumes normality; r = Σ((xi - x̄)(yi - ȳ)) / √(Σ(xi - x̄)²Σ(yi - ȳ)²).
* **Spearman’s rho**: Ordinal or non-normal data; rank-based correlation.
* **Correlation Coefficient (r)**:
  + Range: -1 to +1.
  + |r| = 0: No linear relationship; |r| = 1: Perfect linear relationship.
  + Strength: Weak (< 0.3), Moderate (0.3-0.7), Strong (> 0.7).
  + Direction: Positive (↑x, ↑y), Negative (↑x, ↓y).
* Example: r = -0.85 → strong negative relationship (e.g., hours studied vs. errors made).

**Steps to Generate Rules**:

1. Find frequent itemsets using minimum support threshold (e.g., support ≥ 0.2).
2. Generate rules from itemsets using minimum confidence threshold (e.g., confidence ≥ 0.5).
3. Evaluate rules with lift to assess interestingness (lift > 1 = strong rule).

**Use Cases**:

* **Market Basket Analysis**: Identify products bought together (e.g., {diapers → beer}).
* **Product Placement**: Place related items near each other in stores.
* **Recommendation Systems**: Suggest items (e.g., “Customers who bought this also bought…”).
* **Cross-Selling**: Bundle products for promotions.

**Data Format**:

* Requires binomial form: 1 = item present, 0 = absent.
* Example: Transaction {milk, eggs} → [1, 1, 0] for milk, eggs, bread.

**Metrics**:

* **Support**: Proportion of transactions with itemset.
  + Formula: Support(A) = (# transactions with A) / (total transactions).
  + Example: {bread, peanut butter} in 2/4 transactions → Support = 0.5.
* **Confidence**: Likelihood consequent occurs given antecedent.
  + Formula: Confidence(A → B) = Support(A, B) / Support(A).
  + Example: {peanut butter → eggs}, Support(peanut butter, eggs) = 0.25, Support(peanut butter) = 0.5 → Confidence = 0.25/0.5 = 0.5.

**Module 3.2: Association Rules Analysis**

**Key Definitions**:

* **Association Analysis**: Data mining to find patterns of co-occurring items in transactions (e.g., market basket analysis).
* **Itemset**: Set of items in a transaction (e.g., {milk, bread}).
* **Association Rule**: If {antecedent} → Then {consequent} (e.g., {milk} → {bread}).
* **Antecedent**: “If” part (condition, e.g., {milk}).
* **Consequent**: “Then” part (outcome, e.g., {bread}).
* Lift: strength of a rule vs random co-occurrence
  + Formula: lift (a -> b) = confidence (a -> b)/support(b)
  + Example confidence (peanute butter -> eggs) = 0.5, support eggs (0.75), lift = 0.5/0.75 = 0.67

.**Lift Interpretation**:

**Key Concepts**:

* **Clustering**: Unsupervised learning to group similar observations without labels.
* **k-means**: Partitions data into k clusters by minimizing intra-cluster distances.
* **Centroid**: Mean of all points in a cluster; represents cluster center.
* **k**: Number of clusters, chosen by analyst.
* **Intraclass Homogeneity**: Points within a cluster are similar (minimize within-cluster variance).
* **Interclass Separation**: Clusters are distinct (maximize between-cluster distance).

**Distance Metric**:

* **Euclidean Distance**: Measures similarity; shorter = more similar.
  + Formula: d = √((x2-x1)² + (y2-y1)² + ...).
  + Example: Points A(8, 0), B(4, -3) → d = √((8-4)² + (0-(-3)²) = √(16+9) = 5.
* Other metrics: Manhattan (sum of absolute differences), Cosine (angle-based).

**k-means Algorithm Steps**:

1. Choose k; randomly initialize k centroids.
2. Assign each point to nearest centroid (using Euclidean distance).
3. Recalculate centroids as mean of assigned points.
4. Repeat steps 2-3 until centroids stabilize or max iterations reached.

**Importance of Normalization**:

* Scales variables to equal ranges (e.g., z-scores: (x - μ)/σ).
* Prevents dominance by large-scale variables (e.g., salary in $ vs. age in years).
* **Interpreting Normalized Centroids**:
  + Positive = above mean, negative = below mean, |value| = standard deviations.
  + Example: Salary centroid = 0.391 → 0.391 SDs above mean salary.

**Choosing Optimal k**:

* **Elbow Rule**: Plot within-cluster sum of squares (WSS) vs. k; choose k at “elbow” where WSS drop slows.
* **Silhouette Score**: Measures cluster cohesion/separation (range: -1 to 1; higher = better).
* **Gap Statistic**: Compares WSS to random data WSS; choose k with max gap.
* **Post-Hoc Evaluation**: Assess clusters for interpretability/business relevance.
* Example: Elbow plot shows WSS flattens at k = 3 → likely optimal.

**Handling Categorical Variables**:

* Encode to numeric:
  + **One-hot encoding**: ‘Red’ → [1, 0, 0], ‘Blue’ → [0, 1, 0].
  + **Ordinal encoding**: Low = 1, Medium = 2, High = 3.
* Caution: High dimensionality (many categories) may skew distances; consider feature selection.
* Example: Department = {Sales, HR} → [1, 0] or [0, 1].

**Limitations**:

* Sensitive to outliers and initial centroid placement.
* Assumes spherical clusters; struggles with non-linear shapes.
* Requires predefined k; wrong k may lead to poor clusters.
* May not find meaningful groups in noisy/complex data.

**Interpreting Clusters**:

* Examine centroid values: High/low values indicate defining traits.
* Name clusters based on prominent attributes (e.g., “High Salary, Senior” for cluster with Salary = 1.2, Job Level = 1.4).
* Visualize: Parallel coordinate plots (lines = variables across clusters), scatter plots (points by cluster).
* Example: Cluster with Age = -0.5, Salary = 0.8 → “Young, High Earners.”

**Common Business Uses**:

* **Customer Segmentation**: Group by purchase behavior (e.g., frequent vs. occasional buyers).
* **Targeted Marketing**: Tailor campaigns to cluster traits.
* **Anomaly Detection**: Identify outliers as unusual patterns.
* **Inventory Management**: Cluster products by demand patterns.

**Glossary**:

* **Multicollinearity**: High correlation between predictors, complicating regression.
* **Intraclass Homogeneity**: Similarity within clusters (low intra-cluster distance).
* **Silhouette Score**: Metric for cluster quality (higher = better separation).
* **WSS**: Within-cluster sum of squares; measures cluster compactness.
* **Type I Error**: Rejecting true H0 (false positive).
* Range: 0 to ∞.
* Lift = 1: No association (independence).
* Lift > 1: Positive association (items co-occur more than expected).
* Lift < 1: Negative association (items co-occur less than expected).
* Example: Lift = 3 for {milk, eggs → bread} → 3x more likely to buy bread with milk and eggs.

**Example Calculation**:

* Dataset: {1: milk, eggs, bread}, {2: bread, peanut butter}, {3: peanut butter, eggs, jam}, {4: cereal, peanut butter, bread}.
* Support {bread, peanut butter}: 2/4 = 0.5.
* Confidence {peanut butter → bread}: Support(peanut butter, bread) / Support(peanut butter) = 0.5 / 0.75 = 0.67.
* Lift {peanut butter → bread}: Confidence / Support(bread) = 0.67 / 0.75 ≈ 0.89 (negative association).

**Practice Problem**:

* Dataset: {1: A, B, C}, {2: B, D}, {3: A, B}, {4: C, D}.
* Calculate: Support {A, B}, Confidence {A → B}, Lift {A → B}.
* Answer: Support = 2/4 = 0.5; Confidence = 0.5/0.5 = 1; Lift = 1/(3/4) ≈ 1.33.

**Module 3.3: Clustering Analysis**

**Outliers**:

* Can distort centroids, pulling clusters toward extreme values.
* Detection: Boxplots (beyond 1.5\*IQR), z-scores (>3 SDs).
* Handling: Exclude if unrepresentative, use robust clustering (e.g., k-medoids), or analyze separately.
* Example: Salary $20,000 vs. $500-$2,000 → assess if valid segment or noise.

**CRISP-DM Framework**:

* Steps: 1) Business Understanding, 2) Data Understanding, 3) Data Preparation, 4) Modeling, 5) Evaluation, 6) Deployment.
* Pre-modeling: Define goals, clean data, normalize, encode variables.
* Example: Before clustering, ensure missing values filled and variables scaled.

**Practice Problem**:

* Dataset: Points A(1, 2), B(3, 4), C(2, 3).
* Calculate: Euclidean distance A-B (√((3-1)²+(4-2)²) = √8 ≈ 2.83).
* Interpret: Centroid Salary = -0.7, Age = 1.1 → “Older, Low Earners.”

**Additional Notes & Tools**

**Common Misconceptions**:

* **Correlation**: r ≠ causation; r = 0.9 doesn’t mean x causes y (e.g., ice cream and drowning).
* **Association**: Support is a proportion, not count (e.g., 2/4, not 2).
* **Clustering**: Doesn’t predict outcomes; finds patterns. Clusters may not exist in noisy data.
* **Centroids**: Normalized values reflect relative means (e.g., 0.5 ≠ raw value 0.5).
* **k-means**: Assumes equal-sized, spherical clusters; may fail for irregular shapes.

**Visual Aids** (Descriptions):

* **Correlation Scatter Plot**: x = hours studied, y = test score; positive slope for r = 0.7.
* **Association Rule Diagram**: Arrow from {milk, eggs} → {bread} with support/confidence/lift.
* **Elbow Plot**: WSS vs. k; curve bends at k = 3, suggesting optimal k.
* **Parallel Coordinate Plot**: Lines connect centroid values across variables (e.g., Salary, Age) for clusters.

**Final Tips**:

* Double-check calculations (e.g., support = transactions/total, not raw count).
* Normalize before clustering to avoid variable dominance.
* Use lift > 1 to prioritize strong association rules.
* Validate clusters with business context; meaningless clusters waste resources.
* Practice interpreting p-values, centroids, and lift with real datasets.