**Data Handling**:

1. **How would you handle missing values in a dataset? Describe at least two methods.**

Handling missing values in a dataset is crucial for ensuring the quality and accuracy of data analysis or modelling. Here are two commonly used methods:

**1. Imputation**

**What it is:** Replacing missing values with a substitute value based on the rest of the data.

**Techniques:**

**Mean/Median Imputation:** Replace missing values with the mean (for normally distributed data) or median (for skewed data) of the column.

**Mode Imputation:** Replace with the most frequently occurring value (useful for categorical variables)

**2. Deletion**

**What it is:** Removing rows or columns with missing values.

**Techniques:**

**Listwise Deletion (Complete Case Analysis):** Remove rows with any missing values. Works well if the percentage of missing data is small and the missingness is random.

**Pairwise Deletion:** Exclude rows only for analyses involving variables with missing values. This retains more data but can complicate analysis.

**Column Deletion:** Remove columns with a high percentage of missing values (e.g., over a predefined threshold like 50%).

1. **Explain why it might be necessary to convert data types before performing an analysis.**

Converting data types before analysis is essential for ensuring that the data is interpreted and processed correctly by statistical or machine learning tools. Here are key reasons why it might be necessary:

**1. Data Consistency and Integrity**

* **Why:** Different data types (e.g., numeric, categorical, date-time) have specific characteristics that affect computations and interpretations.
* **Example:** A column containing numerical values stored as strings cannot be used in arithmetic operations or statistical analyses until converted to a numeric type.

**2. Optimizing Operations**

* **Why:** Using appropriate data types can improve computational efficiency and reduce memory usage.
* **Example:** Converting large floating-point numbers to integers (when applicable) or using categorical data types for non-numeric labels can speed up processing.

**3. Ensuring Compatibility with Analytical Tools**

* **Why:** Many analysis tools and libraries expect specific data types for particular operations.
* **Example:** Machine learning models often require numerical input. Therefore, categorical data must be encoded (e.g., one-hot encoding or label encoding) into numeric formats.

**4. Accurate Statistical Analysis**

* **Why:** Misclassified data types can lead to incorrect statistical calculations or interpretations.
* **Example:** Treating a date column as a string instead of a date-time object will prevent time-series analysis or operations like calculating durations.

**5. Avoiding Errors**

* **Why:** Incorrect data types may cause errors or unexpected behavior in scripts or workflows.
* **Example:** Summing a string column containing numbers will concatenate the values rather than compute the total.

**Statistical Analysis**:

**○ What is a T-test, and in what scenarios would you use it? Provide an example based on sales data.**

A **T-test** is a statistical test used to determine whether there is a significant difference between the means of two groups, assuming that the data follows a normal distribution. It helps assess whether observed differences are likely due to chance or reflect a true difference in the population.

**Types of T-tests**

1. **One-Sample T-test:**
   * Compares the mean of a single sample to a known or hypothesized population mean.
   * **Use Case:** Testing whether the average sales of a product in a region match a target value.
2. **Independent (Two-Sample) T-test:**
   * Compares the means of two independent groups.
   * **Use Case:** Comparing the average sales between two different regions.
3. **Paired (Dependent) T-test:**
   * Compares the means of two related groups (e.g., before-and-after scenarios).
   * **Use Case:** Evaluating the impact of a sales training program by comparing sales before and after the training.

**When to Use a T-test**

* To compare group means and assess if the differences are statistically significant.
* When the data is continuous and approximately normally distributed.
* When comparing two groups with relatively equal variances (for two-sample tests).

**Example Using Sales Data**

**Scenario:** A company wants to determine whether a new marketing strategy improved sales. They collect the weekly sales data from two regions: one that implemented the strategy (Region A) and another that did not (Region B).

**Steps:**

1. **Formulate Hypotheses:**
   * Null Hypothesis (H0H\_0H0​): The mean sales in Region A are equal to those in Region B.
   * Alternative Hypothesis (HaH\_aHa​): The mean sales in Region A are significantly different from those in Region B.
2. **Collect Data:**
   * Region A: [1200,1300,1250,1400,1350][1200, 1300, 1250, 1400, 1350][1200,1300,1250,1400,1350]
   * Region B: [1100,1150,1120,1180,1160][1100, 1150, 1120, 1180, 1160][1100,1150,1120,1180,1160]
3. **Perform an Independent T-test:**
   * Compare the means of the two regions using a t-test.
4. **Interpret Results:**
   * If the p-value is less than a significance level (e.g., 0.05), reject H0H\_0H0​, indicating the marketing strategy likely impacted sales.

This approach helps the company make data-driven decisions about the effectiveness of their strategy.

**3) Describe the Chi-square test for independence and explain when it should be used. How would you apply it to test the relationship between shipping mode and customer segment?**

The **Chi-square test for independence** is a statistical test used to determine whether there is a significant association between two categorical variables. It assesses whether the distribution of one variable is independent of the other. This test is particularly useful when dealing with contingency tables, where frequencies are categorized based on two variables.

**When to Use the Chi-Square Test for Independence:**

1. **Data Type**: Both variables should be categorical.
2. **Expected Frequency**: Each cell in the contingency table should have an expected frequency of at least 5 for the test to be valid (though there are adjustments for smaller samples).
3. **Independence**: Observations must be independent (e.g., no repeated measures on the same subject).
4. **Hypothesis**:
   * Null hypothesis (H0H\_0H0​): The two variables are independent.
   * Alternative hypothesis (HaH\_aHa​): The two variables are not independent.

**Application: Relationship Between Shipping Mode and Customer Segment**

Suppose you have data on shipping modes (e.g., Standard, Express, Two-Day) and customer segments (e.g., Consumer, Corporate, Home Office). You want to test if the choice of shipping mode depends on the customer segment.

1. **Create a Contingency Table**: Summarize the data into a contingency table showing the frequency of shipping modes for each customer segment.

| **Shipping Mode/Customer Segment** | **Consumer** | **Corporate** | **Home Office** | **Total** |
| --- | --- | --- | --- | --- |
| Standard | 150 | 80 | 70 | 300 |
| Express | 100 | 50 | 50 | 200 |
| Two-Day | 50 | 40 | 30 | 120 |
| Total | 300 | 170 | 150 | 620 |

1. **State the Hypotheses**:
   * H0H\_0H0​: Shipping mode is independent of customer segment.
   * HaH\_aHa​: Shipping mode is not independent of customer segment.
2. **Calculate the Expected Frequencies**: Use the formula:
3. **Compute the Chi-Square Statistic**: Use the formula:
4. **Determine the Degrees of Freedom**:
5. **Compare with the Critical Value**: Using the Chi-square distribution table and the chosen significance level (α\alphaα, typically 0.05), determine whether to reject H0H\_0H0​.

**Interpret the Results**:

If χ2\chi^2χ2 exceeds the critical value or the ppp-value is less than α\alphaα, reject H0H\_0H0​. This means there is a significant relationship between shipping mode and customer segment.

Otherwise, fail to reject H0H\_0H0​, indicating no evidence of dependence.

**Univariate and Bivariate Analysis**:

What is univariate analysis, and what are its key purposes?

**Key Purposes of Univariate Analysis**

1. **Understanding Data Distribution**:

Provides insights into the spread, shape, and central tendency of a variable.

Helps identify patterns, trends, and anomalies.

**Summarizing the Data**:

Generates summary statistics to describe the variable succinctly (e.g., mean, median, mode, standard deviation, range).

**Detecting Outliers**:

Highlights values that deviate significantly from the rest, which may indicate errors or interesting data points.

**Assessing Data Quality**:

Helps identify missing values, inconsistencies, or incorrect data entries.

**Assisting in Decision-Making**:

Provides essential descriptive information that informs subsequent analyses

**Q) Explain the difference between univariate and bivariate analysis. Provide an example of each.**

**Univariate Analysis Example**

**Scenario:** Analyzing the age distribution of customers in a dataset.

* Compute summary statistics such as mean, median, and standard deviation.
* Visualize the data using a histogram or box plot to understand the distribution of customer ages.

#### **Bivariate Analysis Example**

**Scenario:** Exploring the relationship between customer age (numerical) and total spending (numerical).

* Use a scatter plot to visualize how spending varies with age.
* Compute the correlation coefficient (rrr) to measure the strength and direction of the relationship.
* Fit a linear regression model, if appropriate, to understand how spending is influenced by age.

Alternatively, if one variable is categorical (e.g., customer segment: Consumer, Corporate, Home Office) and the other is numerical (e.g., spending), you could:

* Compare means using box plots for each segment.
* Conduct an ANOVA test to check for significant differences in spending across segments.

### ****Conclusion****

* **Univariate analysis** focuses on describing individual variables, while **bivariate analysis** explores the relationship between two variables.
* They serve different purposes, with univariate analysis often laying the groundwork for more complex analyses like bivariate or multivariate analysis.