**MKTG5883 Exercise 2: Data Preparation and Preprocessing**

**1. Data Cleanup Exercises**

**1.1. Replacing 'male/female' with a Proper Data Type**

The 'Sex' column uses strings ("Male", "Female"). For many analysis techniques, numerical representations are more efficient and suitable.

We'll use binary encoding:

* "Male" will be represented as 1.
* "Female" will be represented as 0.  
  This is a standard practice for representing binary categorical variables.

Binary encoding is efficient for storage and computation. Many machine learning algorithms work directly with numerical data. We *could* also use one-hot encoding (creating separate columns for 'Male' and 'Female'), but it's unnecessary with only two categories.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **EmployeeID** | **Name** | **Sex** | **Age** | **Qualification** |
| 1 | John | 1 | 24 | College |
| 2 | Mary | 0 |  | Bachelor |
| 3 | Alice | 0 | 49 | College |
| 4 | Shara | 0 | 32 | Master |
| 5 | Peter | 1 | 21 | Bachelor |

**1.2. Handling Missing Age Values**

The 'Age' for Mary (EmployeeID 2) is missing.

We'll use mean imputation – replacing the missing value with the average age of the other employees.

* 1. Sum of known ages: 24 + 49 + 32 + 21 = 126
  2. Number of known ages: 4
  3. Average age: 126 / 4 = 31.5

Mean imputation is a simple and common approach. It preserves the overall mean of the 'Age' variable. However, it *does* reduce the variance of the 'Age' column and can potentially bias results if there are many missing values or if the missingness is not random. Other options (like median imputation or using a model to predict the missing age) could be considered in a real-world scenario, but the instructions specify using the average.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **EmployeeID** | **Name** | **Sex** | **Age** | **Qualification** |
| 1 | John | 1 | 24.0 | College |
| 2 | Mary | 0 | 31.5 | Bachelor |
| 3 | Alice | 0 | 49.0 | College |
| 4 | Shara | 0 | 32.0 | Master |
| 5 | Peter | 1 | 21.0 | Bachelor |

**1.3. Representing Categorical Data (Qualification)**

The 'Qualification' column is categorical (College, Bachelor, Master). We need an alternative representation.

Since there are only three categories, we'll use *ordinal encoding*. We'll assign numerical values that reflect a potential order or hierarchy in the qualifications:

* + Bachelor: 1
  + College: 2 (Assuming, no information about college so we assume it better than Bachelor)
  + Master: 3

Ordinal encoding is suitable when there's a clear order or ranking among the categories. It preserves this ordinal relationship, which can be beneficial for some algorithms. We *could* use one-hot encoding here, but it would create three new columns, which might be less efficient than a single ordinal column. The choice depends on the specific analysis or model being used.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **EmployeeID** | **Name** | **Sex** | **Age** | **Qualification** |
| 1 | John | 1 | 24.0 | 2 |
| 2 | Mary | 0 | 31.5 | 1 |
| 3 | Alice | 0 | 49.0 | 2 |
| 4 | Shara | 0 | 32.0 | 3 |
| 5 | Peter | 1 | 21.0 | 1 |

**2. Outliers Detection (Height)**

* **Data:** 130, 132, 138, 153, 133, 110, 132, 129, 135, 134, 136, 133, 133, 134, 135
  1. **Sort the data:** 110, 129, 130, 132, 132, 133, 133, 133, 134, 134, 135, 135, 136, 138, 153
  2. **Find Q2 (Median):** Since there are 15 data points, the median is the 8th value: Q2 = 133
  3. **Find Q1 (Median of the lower half):** The lower half is: 110, 129, 130, 132, 132, 133, 133. Q1 is the 4th value: Q1 = 132
  4. **Find Q3 (Median of the upper half):** The upper half is: 134, 134, 135, 135, 136, 138, 153. Q3 is the 4th value: Q3 = 135
  5. **Calculate IQR:** IQR = Q3 - Q1 = 135 - 132 = 3
  6. **Calculate Lower Bound:** Q1 - 1.5 \* IQR = 132 - 1.5 \* 3 = 132 - 4.5 = 127.5
  7. **Calculate Upper Bound:** Q3 + 1.5 \* IQR = 135 + 1.5 \* 3 = 135 + 4.5 = 139.5
  8. **Identify Outliers:** Any value below 127.5 or above 139.5 is an outlier.
     + 110 is an outlier (below 127.5)
     + 153 is an outlier (above 139.5)
* The IQR method is a robust way to identify outliers, as it's less sensitive to extreme values than methods based on the mean and standard deviation.

**3. Outliers Detection (Weight)**

* **Data:** 37, 40, 39, 51, 41, 30, 39.5, 38.5, 41.5, 37, 39, 38.5, 37, 40, 41
  1. **Sort the data:** 30, 37, 37, 37, 38.5, 38.5, 39, 39, 39.5, 40, 40, 41, 41, 41.5, 51
  2. **Find Q2 (Median):** The median is the 8th value: Q2 = 39
  3. **Find Q1:** The lower half is: 30, 37, 37, 37, 38.5, 38.5, 39. Q1 is the 4th value: Q1 = 37
  4. **Find Q3:** The upper half is: 39.5, 40, 40, 41, 41, 41.5, 51. Q3 is the 4th value: Q3 = 41
  5. **Calculate IQR:** IQR = Q3 - Q1 = 41 - 37 = 4
  6. **Calculate Lower Bound:** Q1 - 1.5 \* IQR = 37 - 1.5 \* 4 = 37 - 6 = 31
  7. **Calculate Upper Bound:** Q3 + 1.5 \* IQR = 41 + 1.5 \* 4 = 41 + 6 = 47
  8. **Identify Outliers:**
     + 30 is an outlier (below 31)
     + 51 is an outlier (above 47)
* Same as with the height data – the IQR method provides a robust outlier detection approach.

**4. Outliers Detection - Box Plot and Standard Deviation**

This section asks for a box plot and uses the standard deviation rule for outlier detection. This is an *alternative* method to the IQR method used above. We'll do both for the height data.

* **Height Data (again):** 130, 132, 138, 153, 133, 110, 132, 129, 135, 134, 136, 133, 133, 134, 135
* **Box Plot:**
  + We already calculated: Min = 110, Q1 = 132, Q2 = 133, Q3 = 135, Max = 153, IQR = 3
  + The box plot would visually represent these values. The box spans from Q1 to Q3, with a line at Q2 (the median). "Whiskers" extend to the furthest data points *within* the 1.5\*IQR range (127.5 and 139.5). The outliers (110 and 153) would be plotted as individual points beyond the whiskers.
* **Standard Deviation Method:**
  + - **Calculate the Mean:** Sum all values and divide by 15: (130+132+...+135) / 15 = 132.67 (approximately)
    - **Calculate the Standard Deviation:** This is a bit more involved. You subtract the mean from each data point, square the result, sum those squared differences, divide by (n-1) where n is the number of data points, and finally take the square root. Using a calculator or software, the standard deviation is approximately 9.31.
    - **Define Outlier Thresholds:** A common rule is to consider data points more than *two* standard deviations (not one, as stated in the original document; two is more standard) from the mean as outliers. Sometimes, three standard deviations are used for a more conservative threshold. We'll use two.
    - Lower Bound: Mean - 2 \* SD = 132.67 - 2 \* 9.31 = 114.05
    - Upper Bound: Mean + 2 \* SD = 132.67 + 2 \* 9.31 = 151.29
  + **Identify Outliers:**
    - * 110 is an outlier (below 114.05)
      * 153 is an outlier (above 151.29)
* **Comparison of Methods:** In this case, both the IQR method and the 2-standard-deviation method identified the same outliers (110 and 153). However, this won't always be the case. The IQR method is generally preferred because it's *resistant* to the influence of extreme values, while the standard deviation method can be heavily influenced by outliers, potentially masking them. If we used a 1-standard-deviation threshold, we'd get many more "outliers," which is usually not desirable.
* **Box Plot Drawing (Conceptual):**  
  You'd draw a number line from roughly 100 to 160. The box would go from 132 to 135, with a line at 133. Whiskers would extend to 129 (the lowest value within the lower bound) and 138 (the highest value within the upper bound). Points at 110 and 153 would be plotted separately.

