**Q 1.7**

**Iris Dataset Data Cleaning Process**

Step 1: Checking for Missing Values in Columns [Features]

|  |  |
| --- | --- |
| Feature | Number of Missing values |
| Sepal Length | 0 |
| Sepal Width | 6 |
| Petal Length | 7 |
| Petal Width | 2 |

Step 2: Checking for Missing Values in Rows [Datapoints]

15 rows had a missing value. No row has more than one missing value so we can use estimation method for the missing values. If any row had more than one missing value, the row would have been dropped.

Step 3: Replacing Values

We replaced missing values in each feature by the species mean value of the feature.

Results

* The correlation between petal width and species increased from 0.955 to 0.957
* The correlation between petal length and species increased from 0.949 to 0.953
* The correlation between sepal width and species increased from -0.324 to -0.365
* The correlation between sepal length and species remained unchanged

**Heart Disease Dataset Cleaning Process**

Step 1: Checking for Missing Values in Columns [Features]

|  |  |
| --- | --- |
| Feature | Number of Missing values |
| cp | 0 |
| slope | 2 |
| thalach | 5 |
| thal | 1 |
| ca | 0 |

Step 2: Checking for Missing Values in Rows [Datapoints]

8 rows had a missing value. No row has more than one missing value so we can use estimation method for the missing values. If any row had more than one missing value, the row would have been dropped.

Step 3: Replacing Values

The approximation method used in replacing missing values was using the mean of the feature grouped by the target for the numeric feature (thalach). For the categorical features (thal and slope), the missing value was replaced with the mode of the feature grouped by the target.

Results

* The correlation between slope and target decreased from 0.419 to 0.416
* The correlation between thalach and target increased from 0.415 to 0.427
* The correlation between ca and target increased from -0.339 to -0.409
* The correlation between thal and target decreased from -0.368 to - 0.256
* The correlation between cp and target decreased from 0.491 to 0.485