# **Week 2 – Data preprocessing & Data Cleaning**

# **Exercise 01: Data Cleanup Exercises**

We want to analyse the dataset related to the field of “human resource”. Here is some of the original dataset we collect:

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

* Replace male/female with proper datatype to facilitate data processing

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

* Fill any missing age values with the average of the employees.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **EmployeeID** | **Name** | **Sex** | **Age** | **Qualification** |  |  |
| 1 | John | 0 | 24 | College |  |  |
| 2 | Mary | 1 | 31.5 | Bachelor |  |  |
| 3 | Alice | 1 | 49 | College |  |  |
| 4 | Shara | 1 | 32 | Master |  |  |
| 5 | Peter | 0 | 21 | Bachelor |  |  |

* Assume that we have only three types of qualifications. Suggest another way represent such kind of caterical data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **EmployeeID** | **Name** | **Sex** | **Age** |  |  |  |
| 1 | John | 0 | 24 |  |  |  |
| 2 | Mary | 1 | 31.5 |  |  |  |
| 3 | Alice | 1 | 49 |  |  |  |
| 4 | Shara | 1 | 32 |  |  |  |
| 5 | Peter | 0 | 21 |  |  |  |

# **Exercise 02: Outliers Detect**

The doctor of a school has measured the height of pupils in a 5th grade class. The result (in cm) is as follows:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 130 | 132 | 138 | 153 | 133 | 110 | 132 | 129 | 135 | 134 | 136 | 133 | 133 | 134 | 135 |

* Which ones are outliers and why?

Sorted = [110, 129, 130, 132, 132, 133, 133, 133, 134, 134, 135, 135, 136, 138, 153]

Median = 133

Q1 = 132

Q3 = 135

Interquartile range = Q3 - Q1 = 3

Boundaries for inner fences = Q3 + (interquartile range \* 1.5) = 139.5

= Q1 - (interquartile range \* 1.5) = 127.5

Outliers: 110, 153

* The weight of those pupils was measured in kg and the results is as follows. Use the same technique to find the outliers.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 37 | 40 | 39 | 51 | 41 | 30 | 39.5 | 38.5 | 41.5 | 37 | 39 | 38.5 | 37 | 40 | 41 |

Sorted = [30, 37, 37, 37, 38.5, 38.5, 39, 39, 39.5, 40, 40, 41, 41, 41.5, 51]

Median = 39

Q1 = 37

Q3 = 41

Interquartile range = Q3 - Q1 = 4

Boundaries for inner fences = Q3 + (interquartile range \* 1.5) = 47

= Q1 - (interquartile range \* 1.5) = 31

Outliers: 30, 51

Hints: Find the Mean (Q2). Q1 is the mean of the left-side data of Q1, Q3 is the mean of the right-side data of Q1. IQR = Q3-Q1.

*Outliers < Q1 – 1.5 \* IQR or > Q3 + 1.5 \* IQR*

* [Optional] We learned from Lecure 1 that data points that lie more than one standard deviation from the mean are considered outliers. Draw the box lot to intuitively understand the outliers as below figure.

