# **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 | 1 | 24 | College |  |  |
| 2 | Mary | 0 |  | Bachelor |  |  |
| 3 | Alice | 0 | 49 | College |  |  |
| 4 | Shara | 0 | 32 | Master |  |  |
| 5 | Peter | 1 | 21 | Bachelor |  |  |

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

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

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

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

# **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?

For the data set in the table the mean=132.77, s= 6.06, 3s=18.18,

z-score of the observation of 153 is (153-132.77)/6.06=3.34,

z-score of 110 is (110-132.77)/6.06=-3.76.

Since the absolute values of z-score of 153 and 110 are more than 3, the height of 153 cm and the height of 110 cm are outliers in the data set.

* 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 |

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*

The mean for the weight is 40 kg. The Standard deviation s=3.02 and 3s=9.06. The median is 40kg while the quartiles at 25% and 75% are 39.125 and 41 respectively. The normal distribution would range from mean ± 3s = [30.93, 49.07]. The box-plot for weight would look like this and shows outliers 30kg and 51kg

* [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.

