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		24	College		
2	Mary			Bachelor		
3	Alice		49	College		
4	Shara		32	Master		
5	Peter		21	Bachelor		

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

EmployeeID	Name	Sex	Age	Qualification		
1	John		24	College		

2	Mary			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**			
1	John		24			
2	Mary					
3	Alice		49			
4	Shara		32			
5	Peter		21			

Outliers Detection

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?

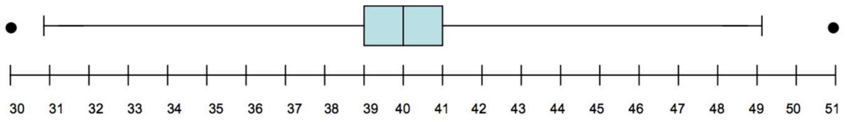
- The weight of those pupils was measured in kg and the results is as follows. Use the same technique to find the outliers.

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*

- [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.


30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51