**Question1.**

**The four potential issues with the dataset are as follows**

a. There are Duplicate Records (A)

b. There is Missing Data (For designation B)

c. The Data columns format is not consistent for DOJ and DOB

d. Primary Key/ Unique identifier is missing.

**Question 2:**

Step 1:

Arrange the data in the ascending order

25

25

30

33

33

35

35

35

35

36

40

41

42

42

99

Note: Q(Denotes Quartile) and IQR denote (Interquartile Range)

Q1 (Quartile1) from the above data is 33

The median of the above data is 35

That is Q2 = 35

Q3 = 41

IQR = Q3 - Q1 = 41 - 33 = 8

**Check for outliers**

Higher Outlier = Q3 + [1.5 \* IQR]

= 41 + [1.5 \* 8]

= 53

Any data that is more than 53 is the outlier. Here in this case 99 is the outlier.

Lower Outlier = Q1 - [1.5 \* IQR]

= 33 - [1.5 \* 8 ] = 33 -12 = 21

There are no lower outliers in the data above.

1. The fix box summary for the above data is as follows

**Size of dataset : 15**

**Median: 35**

**Minimum: 25**

**Maximum: 99**

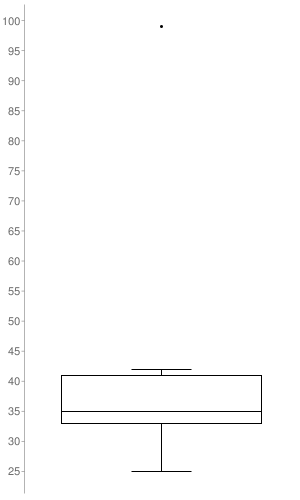
**First quartile: 33**

**Third quartile: 41**

**Interquartile Range: 8**

**Outlier: 99**

1. Draw Box Plot



1. The outliers in this case is 99.

There are no lower outliers and there is only one higher outlier as shown in the data above and that is 99.

1. Effect of outliers on mean and median

Because of the presence of outlier the median does not change. There is no affect on the median. It remains the same in this case 35 or it does not change much.

Because of presence of outlier there is an effect on mean. The mean changes with the introduction of outlier in this case the mean is 39 and if we remove the outlier 99 it changes to 34.

Note: Please refer the Jupyter notebook attached as well.

**Question 3:**

**Attribute 1 Attribute 2 Attribute 3 Attribute 4**

**Object 1 1 1 0 0**

**Object 2 1 0 1 0**

1. **What is the distance between objects if all variables are symmetric ?**

Distance measure for symmetric binary variables

d(Object1, Object2) = r + s / q + r + s + t

      = 1 + 1 / 1 + 1 + 1 + 1

      = 2 / 4 = 1 / 2 = 0.5

1. **What is the distance between objects if all variables are asymmetric ?**

Distance measure for asymmetric binary variables

d(Object 1, Object2) = r + s / q + r + s

        = 1+ 1 / 1+ 1 + 1 = 2 / 3 = 0. 67

**Question 4.**

The two disadvantages of this procedure are mentioned as below with examples

Case 1 : Missing at Random (MAR): Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data  
  
Case 2 : Missing Completely at Random (MCAR): The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.  
  
Case 3 : Missing not at Random (MNAR): Two possible reasons are that the missing value depends on the hypothetical value or missing value is dependent on some other variable’s value.  
  
Example 1: People with high salaries do not want to reveal their income in surveys. Here the results may be highly skewed if we ignore this information.  
  
Example 2: Females do not reveal their age in the data. Here the missing value in age variable is impacted by gender variable.  
  
  
In the first two cases, it is safe to remove the data with missing values depending upon their occurrences, while in the third case removing observations with missing values can produce a bias in the model. So we have to be really careful before removing observations.  
  
Disadvantages of this procedure are as follows  
  
1. The deletion of the data will produce biased parameters and estimates.Reduced Statistical power because it reduces large number of samples and estimates  
will have large standard errors.  
  
2. No chance for imputation to be considered. Poor performance if the percentage of ignored column's rows is high.