amedee

1. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)

1. What is the mean of the data?

x1 = 13+15+16+16+19+20+20+21+22+22+25+25+25+25

x1= 284

x2 = 30+33+33+35+35+35+35+ 36+ 40+45+46+52+70

x2 = 525

*Mean = (x1 +x2)/27*

*Mean* = (284 +525)/27

*Mean = 809/27*

*Mean = 29.96*

1. What is the median?

13 15 16 16 19 20 20 21 22 22 25 25 25 25 30 33 33 35 35 35 35 36 40 45 46 52 70

Median = 25

1. What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc)

Mode = 25 and 35

And our mode is bimodal

1. What is the midrange of the data?

Midrange = (Low + High)/2

Midrange = (13 +70)/2

Midrange = 83/2

Midrange = 41.5

1. Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

Q3 = 30 33 33 35 35 35 35 36 40 45 46 52 70

Q3 = 35

Q1 = 13 15 16 16 19 20 20 21 22 22 25 25 25

Q1= 20

IQR= 35 – 20

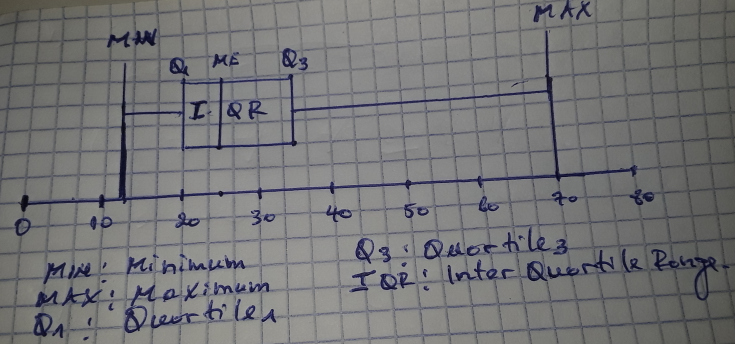
IQR= 15

1. Give the five-number summary of the data.

FIVE NUMBER SUMMARY IS MADE UP WITH:

* Minimum = 13
* First quartile=20
* Median=25
* Third quartile=35
* Maximum=70

1. Show a boxplot of the data.



1. How is a quantile-quantile plot different from a quantile plot?

*Quantile plots directly display the quantiles of a set of values. There is no built-in quantile plot in R, but it is relatively simple to produce one. while Quantile-quantile plots allow us to compare the quantiles of two sets of numbers*

2. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

## Mean or Median Imputation: *When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations.*

## Multivariate Imputation by Chained Equations (MICE)

*MICE assume that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, Bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.*

## 3. Random Forest

*Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple**[decision trees](https://www.datascience.com/blog/random-forests-decision-trees-ensemble-methods" \t "_blank) to estimate missing values and outputs OOB (out of bag) imputation error estimates*

3. you are given this age dataset (13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)

1. Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

*Answer:*

Our Dataset is composed by 27 data so we are going to divide our dataset into 3 equal Bin. Each bin will be composed by 9 elements

* Firs we are going to smoothing this data by equal frequencies beans

Bin1: 13, 15, 16, 16, 19, 20, 20, 21, 22

Bin2: 22, 25, 25, 25, 25, 30, 33, 33, 35

Bin3: 35, 35, 35, 36, 40, 45, 46, 52, 70

* Now let we smooth our data by bin means

Here we will calculate the mean of each single Bin

Bin1: 13, 15, 16, 16, 19, 20, 20, 21, 22

Bin1= (13+15+16+16+19+20+20+21+22)/9

Bin1= 162/9

Bin1= 18, 18, 18, 18, 18, 18, 18, 18,18

Bin2: 22, 25, 25, 25, 25, 30, 33, 33, 35

Bin2 = (22+ 25+ 25+25+25+ 30+33+ 33+35)/9

Bin2= 28.1, 28.1, 28.1, 28.1, 28.1, 28.1, 28.1, 28.1, 28.1

Bin3: 35, 35, 35, 36, 40, 45, 46, 52, 70

Bin3= (35+35+35+36+40+45+46+52+70)/9

Bin3= 43.77, 43.77, 43.77, 43.77, 43.77, 43.77, 43.77, 43.77, 43.77

1. How might you determine outliers in the data?

*One of the simplest methods for detecting outliers is the use of box plots. A box plot is a graphical display for describing the distributions of the data. Box plots use the median and the lower and upper quartiles*

1. What other methods are there for data smoothing?

*Answer:*

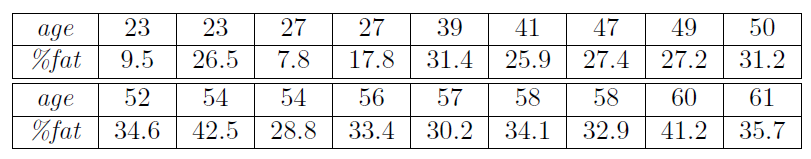
* *Smoothing the data by equal frequency*
* *Smoothing by bin means*
* *Smoothing by bin boundaries*

1. Discuss issues to consider during data integration.

*Answer:*

* [*Varied Information*](https://intellipaat.com/blog/data-integration-process/#_Varied)*: before integrating data, you must check if they are valid and meet all requirements.*
* [*Lack of Manpower*](https://intellipaat.com/blog/data-integration-process/#_Lack)*: all over the world there are a lot of data that can be integrated but the most and big problem Is that the number of skilled people to perform that task is not enough*
* [*Shortage of Storeroom Competence*](https://intellipaat.com/blog/data-integration-process/#_shortage)*:* *There are various challenges to data integration and, among them, the lack of storage is a major issue. If there is no enough space, then it becomes difficult for offering scalability and also elasticity for data. Hence, it stops the growth of data without providing enough space for its proper storage.*
* [*Awful Information*](https://intellipaat.com/blog/data-integration-process/#_Awful)*: Data should be of good quality, or else it affects adversely the data integration process. Databases always have to face serious problems due to impure data features.*
* [*Insights of Data Integration as an Irresistible Attempt*](https://intellipaat.com/blog/data-integration-process/#_Insights)*: The analysis on the information and infrastructure needs a special and foremost promise of capital and also workers.*

4. Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following Result



1. Calculate the mean, median and standard deviation of age and %fat.

Mean age= (23+23+27+27+39+41+47+49+50+52+54+54+56+57+58+58+60+61)/2

Mean age = 46.44444

Mean %fat = 9.5+26.5+7.8+17.8+31.4+25.9+27.4+27.2+31.2+

34.6+42.5+28.8+33.4+30.2+34.1+32.9+41.2+35.7

1. Draw the boxplots for age and %fat.
2. Draw a scatter plot and a q-q plot based on these two variables.
3. Normalize the two variables based on z-score normalization.
4. Calculate the correlation coefficient (Person's product moment coefficient). Are these two variables positively or negatively correlated?

5. Use the two methods below to normalize the following group of data: 200; 300; 400; 600; 1000

1. min-max normalization by setting min = 0 and max = 1
2. z-score normalization

6. Suppose a group of 12 sales price records has been sorted as follows: 5; 10; 11; 13; 15; 35; 50; 55; 72; 92; 204; 215: Partition them into three bins by each of the following methods.

1. equal-frequency partitioning
2. equal-width partitioning