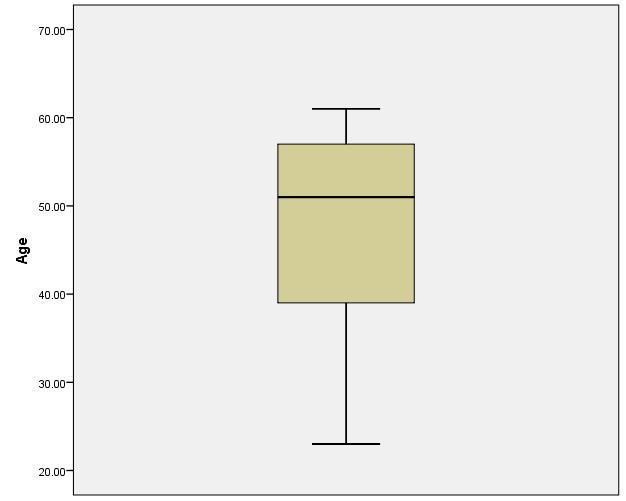
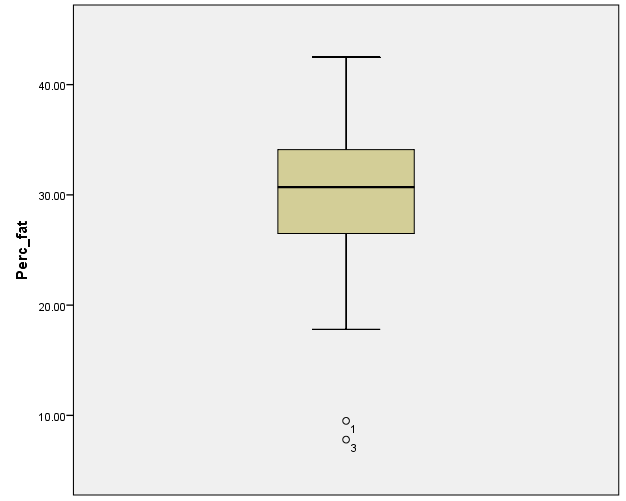
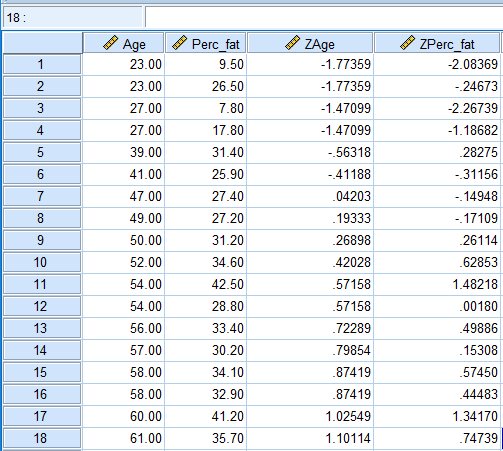
1. Q1
   1. A data warehouse’s primary purpose is for decision making and analysis, while a database’s primary purpose is to record and store data. Databases are to be used for transactions and are in normal formal, and data warehouses are structured to make analysis easier by restricting the data.
   2. OLAP consists of processes that manipulate data in a way that make it easier to data mine. Data Mining happens within the OLAP engine.
   3. A data mart is a subset of a data warehouse that is usually specific to a group of users. There could be many data marts within a corporate data warehouse.
2. Q2
   1. Box plot for age: The boxplot for Age looks to be skewed. The majority of the data comes from ages 40-57, but the minimum is 23.



Boxplot for % fat: The boxplot for %Fat looks to be normally distributed with a mean around 30. It also has 2 outliers which are data points 1 and 3.



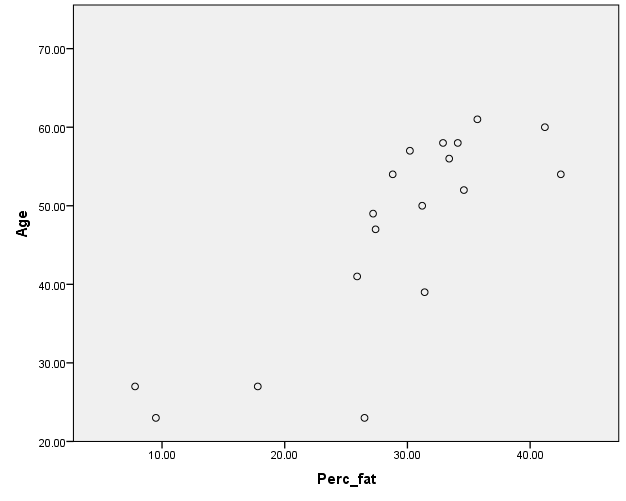


* 1. Min-max normalization is when you change the range to anything you want relative to the original value, usually making it between 0 and 1. This allows you to compare different attributes to each other while using a common scale.

The Z-score normalization tells you how many standard deviations you are from the mean. Not only will this normalize the data, but it will help you detect outliers.

Normalization by decimal scaling is when you divide the numbers by the power of 10, so that all values are less than 1. This will result in all data points being a decimal, allowing you to compare attributes effectively.

* 1. There appears to be a positively correlated relationship between age and percent fat. As age increases, % fat also increase.



|  |  |  |  |
| --- | --- | --- | --- |
| **Correlations** | | | |
|  | | Age | Perc\_fat |
| Age | Pearson Correlation | 1 | .818\*\* |
| Sig. (2-tailed) |  | .000 |
| N | 18 | 18 |
| Perc\_fat | Pearson Correlation | .818\*\* | 1 |
| Sig. (2-tailed) | .000 |  |
| N | 18 | 18 |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). | | | |

Percent fat and age are positively correlated based on a pearson correlation co-efficient of 0.818.

1. Q3:

Bin 1: 5,10,11

Bin 2: 13, 15, 35

Bin3: 50, 55, 72

Bin 4: 92, 204, 215

Bin 1: 5, 10, 11, 13, 15, 35, 50, 55

Bin 2: 72, 92

Bin 3: 204, 215

1. Q4
   1. One way to deal with missing values for some variables is to fill the variable in with the mean of the variable based on the existing values. Another way to deal with missing values which is usually better is to fill it in with the mean of the class it belongs to within each variable. You can also simply just remove the data points that have missing values.
   2. You can use the means of each class to fill in the missing values.
   3. One issue that could come up during data integration is a mis-match of units. For instance, one data set could have units in inches while the other data set could have units in feet.