Problem 1:

a. **What is the difference between data warehouse and database?**

The difference between data warehouse and database is in data warehouse consist of a repository of information which is collected from various sources. The information stored in a data warehouse is under one unified schema. A database contains interrelated data which is managed and accessed by software programs which represents the data in its current status.

b. **What is the difference between data mining and OLAP?**

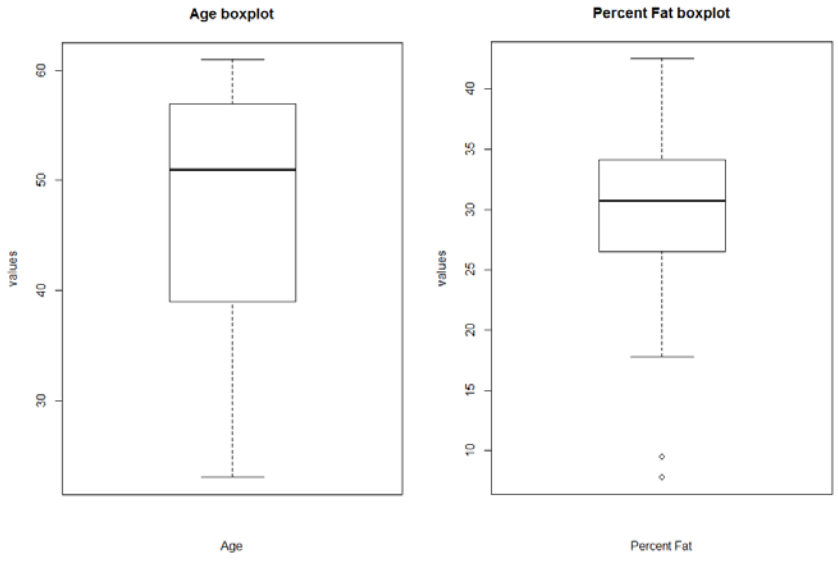
Even though data mining and OLAP are two strategies to solve analytical problems, data mining seeks to discover intriguing patterns from the data. In OLAP the data is summarized and based on the summaries then forecasts are made.

c. **What is the difference between data marts and data warehouse?**

The data marts are part of data warehouse model. A data mart is focused on a specific group or team data. The data mart's scope is quite defined and tailored to that related subject. Whereas the data warehouse encompasses an entire enterprise.

Problem 2:

a. **Draw the box-plots for age and % fat. Explain what you can tell from this visualization of the distribution of the data.**



For the Age, the median is closer to the third quartile. There appears to be several minimum values for the age distributions. For the percent fat boxplot, there appears to be two outliers and median is closer the third quartile.

b. Normalize the two attributes based on z-score normalization. Include an image showing the data table with this done.

```
> z_percent.fat
              [,1]
 [1,] -2.083694688
 [2,] -0.246729622
 [3,] -2.267391195
 [4,] -1.186823509
 [5,]  0.282748545
 [6,] -0.311563683
 [7,] -0.149478530
 [8,] -0.171089884
 [9,]  0.261137191
[10,]  0.628530204
[11,]  1.482178676
[12,]  0.001800946
[13,]  0.498862082
[14,]  0.153080422
[15,]  0.574501820
[16,]  0.444833697
[17,]  1.341704877
[18,]  0.747392650
```

```
> z_age
              [,1]
 [1,] -1.77359187
 [2,] -1.77359187
 [3,] -1.47098852
 [4,] -1.47098852
 [5,] -0.56317846
 [6,] -0.41187679
 [7,]  0.04202824
 [8,]  0.19332992
 [9,]  0.26898076
[10,]  0.42028243
[11,]  0.57158411
[12,]  0.57158411
[13,]  0.72288579
[14,]  0.79853662
[15,]  0.87418746
[16,]  0.87418746
[17,]  1.02548914
[18,]  1.10113998
```

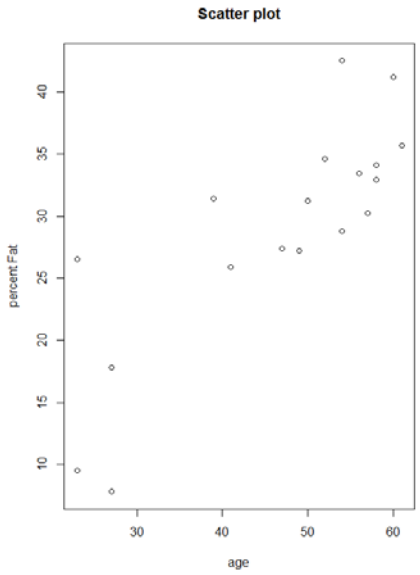
c. Regardless of the original ranges of the variables, normalization techniques transform the data into new ranges that allow to compare and use variables on the same scales. What are the value ranges of the following normalization methods applied to this data? Explain your answer by explaining how the methods work on data in general.

I. **Min-Max normalization (use default target interval 0 to 1)**
The min-max normalization range would be the [new_min, new_max]. Therefore the min would be 0 and the max is 1.

II. **Z-score normalization**
The range for the z-score normalization is [(oldmin-mean)/standard_dev, (oldmax-mean)/standard_dev]. It should be noted that the range is possible for negative infinity to positive infinity.

III. **Normalization by decimal scaling**
The range for this normalization would be (-1.0,1.0).

d. Draw a scatter-plot based on the two variables and visually interpret the relationship between the two variables.



There appears to be a positive relationship between the two variables.

e. **Correlation is useful when integrating or cleaning data to see if two variables are so strongly correlated that they should be checked to see if they duplicate information. Get the full covariance and correlation matrix giving the relationships between all pairs of variables, even though there are only two. Are these two variables positively or negatively correlated?**

Below are the Covariance and Correlation Matrices, there appears to be a positive correlation between the two variables.

Covariance Matrix

```
> cov(M)
                Age Percent_fat
Age         174.7320   100.01961
Percent_fat 100.0196    85.64382
```

Correlation Matrix

```
> round(res, 2)
             Age Percent_fat
Age         1.00        0.82
Percent_fat 0.82        1.00
```

Problem 3:

This problem is an example of data preprocessing needed in a data mining process. 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 bins by each of the following methods. Show which values are in which bins. Then smooth the data using the bins and show the new set of smoothed values. Explain how each type of smoothing affect the data and the ways they are different.

a. Equal-depth partitioning with 3 values per bin.

The width for each interval is (215-5)/3 =70.

Bin 1 = 5, 10, 11, 13, 15, 35, 50, 55, 72
Bin 2 = 92
Bin 3 = 204, 215

Smoothing by bin means:

Bin 1 = 29.5, 29.5, 29.5, 29.5, 29.5, 29.5, 29.5, 29.5, 29.5
Bin 2 = 92
Bin 3 = 209.5, 209.5

Smoothing by bin boundaries:

Bin 1 = 5, 5, 5, 5, 5, 5, 72, 72, 72
Bin 2 = 92
Bin 3 = 204, 215

b. Equal-width partitioning with 3 bins.

Bin 1 = 5, 10, 11, 13
Bin 2 = 15, 35, 50, 55
Bin 3 = 72, 92, 204,215

Smoothing by bin means:

Bin 1 = 9.75, 9.75, 9.75, 9.75
Bin 2 = 38.75, 38.75, 38.75, 38.75
Bin 3 = 145.75, 145.75, 145.75, 145.75

Smoothing by bin boundaries:

Bin 1 = 5, 13, 13, 13
Bin 2 = 15, 15, 55, 55
Bin 3 = 72, 72, 215, 215

Overall goal of smoothing is to remove noise from the data. To do this we can we can utilize several techniques such as binning, regression, and clustering. In this problem we look at binning. The equal-width range is uniform. It is quite obvious that outliers are dominate and likewise the skewness of the data. When smoothing the data for equal width, the greater values have more effect on the smoothing. Likewise, the bins with equal width will have constant values which can be seen in the data. For equal depth (frequency), when smoothing there is a lopsided amount of values for bin 1. Not sure, if the scaling for will be good. (NOTE: created R code just to bin for equal depth and width not to smooth)

Problem 4:

Answer the following questions about the data cleaning and integration process:

a. In real-world data, there are often rows that have missing values for some variables. Describe two methods for dealing with this problem.

When the data has missing values for some variables, there are a few strategies to deal with this problem. For instance, you can ignore the tuple, fill in the missing value (manually), impart a global constant (unknown or minus infinity) on the data. However, one can use regression, Bayesian inference, or decision trees to fill in the missing values in the data. Users have the option to use the central tendency measures such as mean or median to fill in the missing values.

b. If we have class labels for our data, how can we use them to help get better estimates when filling in missing values?

If we have class labels for our data, when implementing classification when filling in missing values. However this is not effective when the percentage of the missing values per attribute varies considerably.

c. Describe two issues that may come up during data integration.

The two issues that may come up during data integration are semantic heterogeneity and data structure. Because the data will be combined from multiple sources there is a high likelihood that schema integration and object matching with pose a great challenge. The data scientists must be mindful of the attributes matching to the correct items as well as attribute types and consistency. In data integration, redundancy is additional issue to deal with especially when one attribute can be derived from another. Therefore, using correlation analysis will help assess whether redundancies are present in the data. Depending on the type of data one has, we use chi-square for nominal data and correlation coefficient and covariance for numeric data. Also with data integration, there are some issues of with how there are different representation of the data values. For instance, the scales of the data or how the data is encoded can cause value conflicts.

Bonus Problem:

We discussed how a clustering of data can be used to smooth data, so let's consider if it could be used for repairing missing data. We discussed how class labels can be used to improve the process of filling in missing values (and you wrote about it in 3b), and we discussed how a clustering result can be used similarly to class labels. Can we cluster data and use the clustering to fill in missing values? If so, how? If not, what problem would we encounter?

To utilize the clustering technique in order to fill in missing values will result is labyrinth experience which would not be recommended. When clustering algorithms are initialized they ignore missing values so if one's objective is to fill in missing values using clustering this would defeat its purpose. The clustering techniques would mess up the distance calculations and eigenvalues