# IS 467

# Assignment 2

**Submission Instructions**

1. Save your solutions with **clearly marked problem numbers**, clear and succinct writing, all software **output** into a single PDF or Word file.
2. Submit your file online at the course website at **http://d2l.depaul.edu** and double-check it.
3. Keep a copy of all your submissions!
4. If you have questions about the homework, email me BEFORE the deadline.

**Problem 1 (15%):**
Answer each of the following questions with a few sentences:

a. What is the difference between *data warehouse* and *database*?
*b.* What is the difference between *data mining* and *OLAP*?
c. What is the difference between *data marts* and *data warehouse*?

**Problem 2 (35% points):** This problem is an example of data preprocessing needed in a data mining process.

Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Age | 23 | 23 | 27 | 27 | 39 | 41 | 47 | 49 | 50 |
| %fat | 9.5 | 26.5 | 7.8 | 17.8 | 31.4 | 25.9 | 27.4 | 27.2 | 31.2 |
| Age | 52 | 54 | 54 | 56 | 57 | 58 | 58 | 60 | 61 |
| %fat | 34.6 | 42.5 | 28.8 | 33.4 | 30.2 | 34.1 | 32.9 | 41.2 | 35.7 |

Use the data view to enter this into an SPSS table. Then do the following –

a. (7%) Draw the box-plots for age and % fat. Explain what you can tell from this visualization of the distribution of the data.

b. (7%) Normalize the two attributes based on z-score normalization. Include an image showing the data table with this done.

c. (7%) 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)

   ii. Z-score normalization

   iii. Normalization by decimal scaling.

d. (7%) Draw a scatter-plot based on the two variables and visually interpret the relationship between the two variables.

e. (7%) 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?

**Problem 3 (25%):** 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. (10%) equal-depth partitioning with 3 values per bin

b. (15%) equal-width partitioning with 3 bins

**Problem 4 (25%):** Answer the following questions about the data cleaning and integration process:

a. (10%) In real-world data, there are often rows that have missing values for some variables. Describe two methods for dealing with this problem.

b. (5%) If we have class labels for our data, how can we use them to help get better estimates when filling in missing values?

c. (10%) Describe two issues that may come up during data integration.

**Bonus Problem (15%):**

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 4b), 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?