Name: Amy Aumpansub IS 467
Due: May 8, 2018

Assignment 2

Problem 1

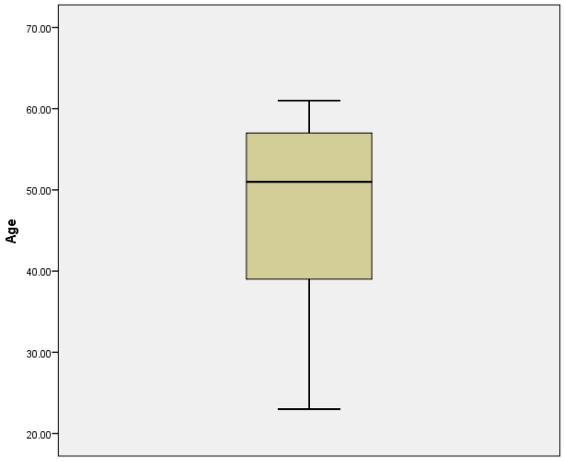
a. Data warehouse focuses on analysis. It summarizes data, collects data from multiple sources and stores data over historical time. It keeps data de-normalized and allows data redundancy. Whereas, database is performance orientated focusing on accuracy and designed for transaction processing. Database primarily manages current data and keeps data normalized, thus it has no data redundancy or no duplication.

b. OLAP is a data aggregation and summarization tool providing multidimensional data. OLAP is aimed to summarize and simplify data to support analysis. Whereas, data mining operates at a detail level not a summary level and is aimed to support knowledge discovery of hidden patterns. Compared to OLAP, data mining has a boarder view because in addition to data summarization, data mining can perform other tasks such as prediction, classification, etc.

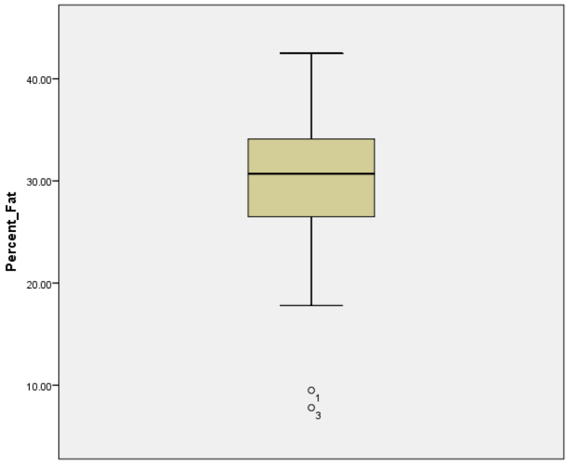
c. Data mart and data warehouse differs in their scope and data sources. Data mart is a subset of corporate-wide data. Data mart focuses on a specific scope and is designed for a selected group of users, thus it mainly integrates data from a selected source. Whereas data warehouse has an enterprise-wide depth, covering multiple subject areas which operates to integrate data from multiple sources.

Problem 2

a. Draw the box-plots for age and % fat



Age

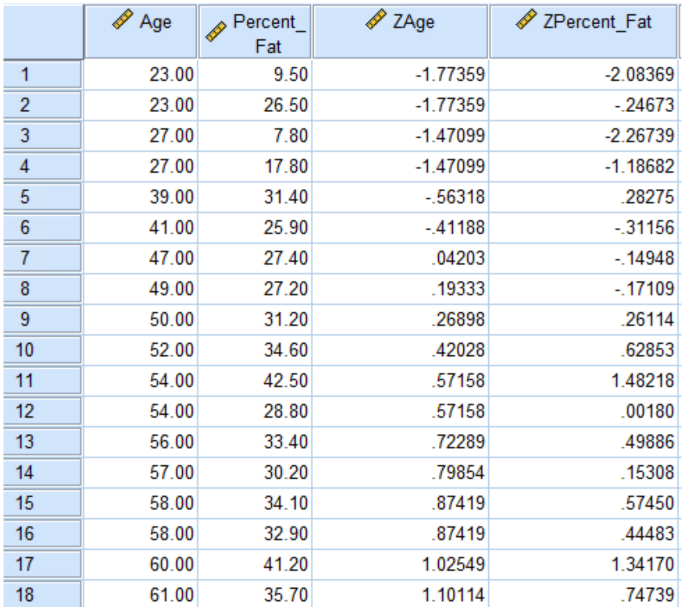


% Fat

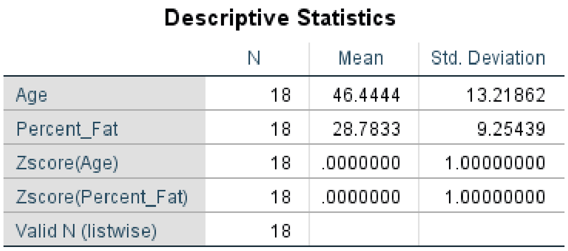
For Age, the distribution is not symmetric and skewed left. The “Age” box plot shows the minimum age of 23, median age of 51, and maximum age of 61. The plot shows no potential outliers.

For % Fat, the distribution is normal, but data contains two potential outliers (observation number 1 and 3). The “%Fat” box plot shows the minimum % Fat of 7.8, median % Fat of 30.7, and maximum % Fat of 42.50. Compared to %Fat, the “Age” distribution has a larger interquartile.

b. Z-score normalization



	Age	Percent_Fat	ZAge	ZPercent_Fat
1	23.00	9.50	-1.77359	-2.08369
2	23.00	26.50	-1.77359	-.24673
3	27.00	7.80	-1.47099	-2.26739
4	27.00	17.80	-1.47099	-1.18682
5	39.00	31.40	-.56318	.28275
6	41.00	25.90	-.41188	-.31156
7	47.00	27.40	.04203	-.14948
8	49.00	27.20	.19333	-.17109
9	50.00	31.20	.26898	.26114
10	52.00	34.60	.42028	.62853
11	54.00	42.50	.57158	1.48218
12	54.00	28.80	.57158	.00180
13	56.00	33.40	.72289	.49886
14	57.00	30.20	.79854	.15308
15	58.00	34.10	.87419	.57450
16	58.00	32.90	.87419	.44483
17	60.00	41.20	1.02549	1.34170
18	61.00	35.70	1.10114	.74739



Descriptive Statistics

	N	Mean	Std. Deviation
Age	18	46.4444	13.21862
Percent_Fat	18	28.7833	9.25439
Zscore(Age)	18	.0000000	1.00000000
Zscore(Percent_Fat)	18	.0000000	1.00000000
Valid N (listwise)	18		

Data are normalized based on z-score. The descriptive statistic shows the normalized variables (ZAge and ZPercent_Fat) have a 0 mean and standard deviation of 1.

c.

i. Min-max normalization

The range is **[new min, new max],** so the new range for both age and % fat is **[0, 1].**
This method normalizes the original data by performing a linear transformation. It uses the original maximum and minimum values and the new maximum and minimum values to transform data.
For this problem, Age $\longrightarrow$ $v' = \frac{V_{age} - 23}{61-23}(1-0)+0$ %Fat $\longrightarrow$ $v' = \frac{V_{fat} - 7.8}{42.5-7.8}(1-0)+0$

ii. Z-score normalization

The range is **[(old min − mean)/stddev, (old max−mean)/stddev] or [- ∞,∞]** for all possible datasets.
The new range of age is **[-1.77359, 1.10114].** The new range of % fat is **[-2.26739, 1.48218].**

The z-score normalization transformed the original data based on its original mean and standard deviation. The transformed data falls within a smaller range and has a normal distribution with 0 mean and a standard deviation of 1.
The normalized value (v') is: Age $\longrightarrow$ $v' = \frac{V_{age} - 46.4}{13.2}$ %Fat $\longrightarrow$ $v' = \frac{V_{fat} - 28.7}{9.25}$

iii. Normalization by decimal scaling

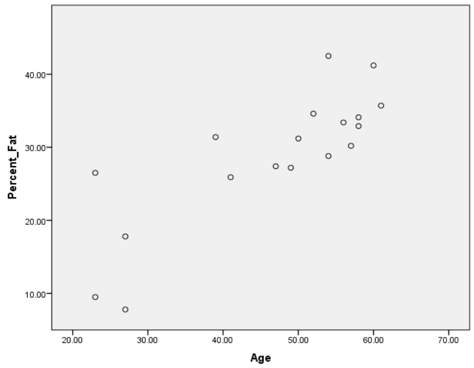
The range is **[-1,1].** the new range of age is **[0.23, 0.61].** The new range of % fat is **[0.078, 0.425].**

This method normalizes the original data by moving the decimal point of the original values, so the new range is smaller. The maximum absolute value after normalization is less than 1. The number of decimal points to be moved (j) is computed from the maximum absolute value, so in this problem j = 2 because max│Age│= 61 and max│%fat│= 42.5.
The normalized value is: Age $\longrightarrow$ $v' = \frac{V_{age}}{10^2}$ %Fat $\longrightarrow$ $v' = \frac{V_{fat}}{10^2}$

Name: Amy Aumpansub IS 467
Due: May 8, 2018

d.

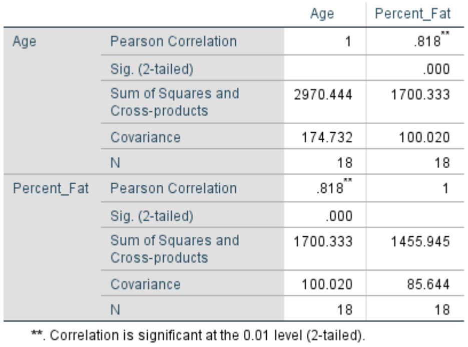
A scatter-plot



The plot shows a linear trend between two variables indicating that the linear relationship between age and % fat is positive and strong, in which the % fat increases as age increases.

e.

Covariance and Correlation matrix



		Age	Percent_Fat
Age	Pearson Correlation	1	.818**
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	2970.444	1700.333
	Covariance	174.732	100.020
	N	18	18
Percent_Fat	Pearson Correlation	.818**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	1700.333	1455.945
	Covariance	100.020	85.644
	N	18	18

**. Correlation is significant at the 0.01 level (2-tailed).

The matrix shows the correlation of 0.818 indicating that age and % fat are positively and strongly correlated. As age increases, % fat also increases. The covariance between them is 100.02 confirms that there is a positive relationship. The matrix also displays the variance of age of 174.73 and the variance of % fat of 85.64.

Problem 3

a. equal-depth partitioning

Partition into bins:

Bin 1: 5, 10, 11	Bin 2: 13, 15, 35	Bin 3: 50, 55, 72	Bin 4: 92, 204, 215

Smoothing data by bin means:

Bin 1: 8.67, 8.67, 8.67	Bin 2: 21, 21, 21	Bin 3: 59, 59, 59	Bin 4: 170.3, 170.3, 170.3

Smoothing data by bin medians:

Bin 1: 10, 10, 10	Bin 2: 15, 15, 15	Bin 3: 55, 55, 55	Bin 4: 204, 204, 204

Smoothing data by bin boundaries:

Bin 1: 5, 11, 11	Bin 2: 13, 13, 35	Bin 3: 50, 50, 72	Bin 4: 92, 215, 215

b. equal-width partitioning

Partition into bins: (215-5)/3 = 70

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

Smoothing data 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 data by bin medians:

Bin 1: 15, 15, 15, 15, 15, 15 15, 15, 15	Bin 2: 92	Bin 3: 209.5, 209.5

Smoothing data by bin boundaries:

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

Smoothing data by using binning allows the noise removal, then we smooth data by consulting each neighborhood. Smoothing by bin means and medians replaces each value by the bin means and medians, while smoothing by bin boundaries replaces each bin value by the closest boundary value.

For the equal-depth partitioning, each bin contains the approximately same number of samples. It takes the data distribution into account which is good for data scaling, but hard to handle categorical attributes.

Whereas, equal-width partitioning has the constant interval range of values in each bin. This method cannot handle skewed data very well and the outliers can dominate presentation.

Name: Amy Aumpansub IS 467
Due: May 8, 2018

Problem 4

a.

First method, we can use a global constant to fill in the missing value by replacing missing values by the same label such as “unknown.” Second method, we can use a measure of central tendency such as mean and median of each attribute to replace the missing values. For example, we can use the mean of revenue to replace the missing value of revenue. Third method, we can use the most probable value determined by regression, decision tree, Bayesian inference to replace the missing values.

b.

If data have class labels, we can fill in the missing values by using the attribute mean or mean for all samples within the same class as the given tuple. For example, the customer data has a credit risk class. We can fill in the missing value of income by using the mean of income for the customers in the same credit risk class. This method provides a better estimation of missing values.

c.

Schema integration can be an issue. In this case, the metadata from different sources should be integrated to match the equivalent the real-world entities, referred as the entity identification problem. It is a problem of object matching from multiple sources that correspond to equivalent the real-world entities.

Second issue is data redundancy. For example, the attribute can be redundant if it can be derived from another attribute. Inconsistencies in attribute or dimension naming can also cause the data redundancy.

Bonus Problem

From my perspective, clustering can be used to identify the missing values in which the observations with missing values will be grouped into the same cluster. Then we can implement the similar steps as we implement in the class label data. After we identify the missing values, we calculate the attribute mean of remaining samples within the same cluster label. Then we fill in the missing value with the attribute mean which corresponds to each cluster.

The main issue is that if we do not handle missing values properly, our cluster analysis will be biased because all missing values will be grouped into one cluster.