# Solutions to Chapter 1
# AN INTRODUCTION TO DATA MINING
**Prepared by James Cunningham, Graduate Assistant**

1. **Refer to the Bank of America example early in the chapter. Which data mining task or tasks are implied in identifying "the type of marketing approach for a particular customer, based on customer's individual profile"? Which tasks are not explicitly relevant?**

   Relevant tasks include the following:
   - Description
   - Classification
   - Clustering
   - Associating

   Non-relevant tasks are:
   - Estimation
   - Prediction

2. **For each of the following, identify the relevant data mining task(s):**

   a. **The Boston Celtics would like to approximate how many points their next opponent will score against them.**

      Estimation: estimating the number of points (numeric target).

   b. **A military intelligence officer is interested in learning about the respective proportions of Sunnis and Shias in a particular strategic region.**

      Description: exploratory data analysis finds similarities and differences between the Sunni and Shias proportions.

   c. **A NORAD defense computer must decide immediately whether a blip on the radar is a flock of geese or an incoming nuclear missile.**

      Classification: a trained model detects incoming missiles assigns the blip on the radar screen (unclassified record) as being either a "missile" or "not missile" (categorical target); Estimation: an estimated numeric value may indicate the blip as an incoming missile.

**d. A political strategist is seeking the best groups to canvass for donations in a county.**

Description: relevant patterns describe the characteristics of one or more groups are located in the county; Clustering: examine the profile of each homogeneous group derived from a particular county's population; Association: discover interesting rules pertaining to a large proportion of the population.

**e. A Homeland Security official would like to determine whether a certain sequence of financial and residence moves implies a tendency to terrorist acts.**

Description: the sequences of financial and residential moves (patterns) may suggest a tendency (explanation) for terrorist activities; Classification: build a model to classify behavior as "suspicious"; Estimation: the model generates a numeric score indicating a propensity for committing terrorist acts.

**f. A Wall Street analyst has been asked to find out the expected change in stock price for a set of companies with similar price/earnings ratios.**

Estimation: the expected change in stock price (numeric target) using the price/earnings ratio for a similar set of companies (predictors); Prediction: applied when results expected to predict future price.

**3. For each of the following meetings, explain which phase in the CRISP-DM process is represented:**

**a. Managers want to know by next week whether deployment will take place. Therefore, analysts meet to discuss how useful and accurate their model is.**

The Evaluation Phase determines whether the data mining model achieves the objectives established in the first phase.

**b. The data mining project manager meets with the data warehousing manager to discuss how the data will be collected.**

Although the data warehouse is identified as a resource during the Business Understanding Phase, the actual data collection takes place during the Data Understanding Phase.

**c. The data mining consultant meets with the Vice President for Marketing, who says that he would like to move forward with customer relationship management.**

The primary objectives of the business are stated as part of the Business Understanding Phase.

**d. The data mining project manager meets with the production line supervisor, to discuss implementation of changes and improvements.**

The requirements of a data mining technique used during the Modeling Phase may cause the process to loop back to the Data Preparation Phase, with the goal of improving data quality. The Evaluation Phase determines whether specific improvements or process changes are required to ensure that all important aspects of the business are accounted for.

**e. The analysts meet to discuss whether the neural network or decision tree models should be applied.**

During the Modeling Phase one or more modeling techniques are chosen.

**4. Discuss the need for human direction of data mining. Describe the possible consequences of relying on completely automatic data analysis tools.**

The case studies emphasize the need for human involvement during every phase of the data mining process. For example, data mining initiatives using legacy database systems should not underestimate the time or importance required from domain experts to interpret the data. Taking shortcuts during this initial phase leads to potentially costly, inaccurate results in subsequent phases.

**5. CRISP-DM is not the only standard process for data mining. Research an alternative methodology. (Hint: SEMMA, from the SAS Institute.) Discuss the similarities and differences with CRISP-DM.**

SEMMA is an acronym representing the core data mining processes: sample, explore, modify, model, and assess. As compared to CRISP-DM, SEMMA places emphasis on the model development process of data mining and therefore does not contain a Business Understanding Phase or a Deployment Phase; however, it does describe the importance of having clear business objectives and using quality data sources for modeling.

Both processes are iterative and may loop back to other process steps as new information is learned or data mining requirements change. Also, both methods emphasize the use of an adaptive process. The following table shows how the phases of the two processes correspond to one another:

| SEMMA | CRISP-DM |
|---|---|
| N/A | Business Understanding Phase |
| Sample | N/A[1] |
| Explore | Data Understanding Phase |
| Modify | Data Preparation Phase |
| Model | Modeling Phase |
| Assess | Evaluation Phase |
| N/A | Deployment Phase |

**Table 5.1. SEMMA vs CRISP-DM**

[1] Although Sample does not correspond to a specific CRISP-DM phase, it often occurs during the Data Understanding, Data Preparation, and Modeling Phases.

# Solutions to Chapter 2
# DATA PREPROCESSING

**Prepared by James Cunningham, Graduate Assistant**

**1. Describe the possible negative effects of proceeding directly to mine data that has not been preprocessed.**

Neglecting to preprocess the data adequately before data modeling begins will likely produce data models that are unreliable and whose results should be considered dubious as best. Performing data cleaning and data transformation during the data preparation phase is absolutely necessary for successful data mining efforts.

For example, suppose you are analyzing a data set that includes a person's Age and Date_of_Birth attributes, and you want to calculate the average Age. Now, if 5% of the records contain a value of 0 for Age, the mean value would be very misleading and inaccurate. One solution to this problem would be to derive Age for the zero-based records based on information contained in the Date_of_Birth variable. Now, the mean value for Age is more representative of those persons in the data set.

**2. Refer to the income attribute of the five customers in Table 2.1, before preprocessing.**

**a. Find the mean income before preprocessing.**

The mean value for Income before preprocessing is 38,999.80 and is derived by the possible inclusion of Income values -40,000 (erroneous) and 100,000 (possible outlier).

**b. What does this number actually mean?**

In this case the mean value has little meaning because we are combining real data values with erroneous values.

**c. Now, calculate the mean income for the three values left after preprocessing. Does this value have a meaning?**

However, the mean value for Income produced by values 75,000, 50,000, and 10,000 (9,999 rounded to nearest 5,000) is 45,000. The latter value is certainly more representative of the true mean for Income, now that the records containing questionable values have been excluded.

**3. Explain why zip codes should be considered text variables rather than numeric.**

Zip codes should be considered text variables because they cannot be quantified on any numeric scale. Even their order has no numerical significance.

**4. What is an outlier? Why do we need to treat outliers carefully?**

Consider a set of numerical observations and the center of this observation set. An outlier is an observation that lies much farther away from the center than the majority of the other observations in the set.

We must treat outliers carefully because they can cause us to misrepresent the true center of an observation set incorrectly if they lie significantly farther away from the other observations in the set.

**5. Explain why a birthdate variable would be preferred to an age variable in a database.**

A birthdate variable is preferable to an age variable in a database because (1) one can always derive age from birthdate by taking the difference from the current date, and (2) age is relative to the current date only and would need to be updated continuously over time in order to remain accurate.

**6. True or false: All things being equal, more information is almost always better.**

The answer is true. In general, more information is almost always better. The more information we have to work with, the more insight into the underlying relationships of a particular domain of discourse we can glean from it.

**7. Explain why it is not recommended, as a strategy for dealing with missing data, to simply omit the records or fields with missing values from the analysis.**

It is not recommended to omit records or fields from an analysis simply because they have missing values. The rationale for this recommendation is that omitting these fields and records may cause us to lose valuable insight into the underlying relationships that we may have gleaned from the partial information that we do have.

**8. Which of the four methods for handling missing data would tend to lead to an underestimate of the spread (e.g., standard deviation) of the variable? What are some benefits to this method?**

Replacing a missing value by the attribute value's mean artificially reduces the measure of spread for that particular attribute. Although the mean value is not necessarily a typical value, for some data sets this form of substitution may work well. Specifically, the effectiveness of this technique depends on the size of the variation of the underlying population. In other words, the technique works well for populations having small variations, and works less effectively for populations having larger variations.

Several benefits to leveraging this method include (1) ease of implementation (i.e. only one value to impute), (2) preservation of the standard error (i.e. no additional residual error is introduced).

**9. What are some of the benefits and drawbacks for the method for handling missing data that chooses values at random from the variable distribution?**

By using the data values randomly generated from the variable distribution, the measures of center and spread are most likely to remain similar to the original; however, there is a chance that the resulting records may not make intuitive sense.

**10. Of the four methods for handling missing data, which method is preferred?**

Having the analyst choose a constant to replace missing values based on specific domain knowledge is overall, probably the most conservative choice. If missing values are replaced with a flag such as "missing" or "unknown", in many situations those records would ultimately be excluded from the modeling process; that is, all remaining valid, potentially important, values contained in those records would not be included in the data model.

**11. Make up a classification scheme which is inherently flawed, and would lead to misclassification, as we find in Table 2.2. For example, classes of items bought in a grocery store.**

| Breakfast | Count |
|---|---|
| Cold Cereals | 72 |
| Sugar Smacks | 1 |
| Cheerios | 2 |
| Hot Cereals | 28 |
| Cream of Wheat | 3 |

**Table 11.1. Flawed classification scheme**

Using the table above, the "Breakfast" categorical attribute contains 5 apparent classes. However, upon further inspection the classes are discovered to be inconsistent. For example, both "Sugar Smacks" and "Cheerios" are cold cereals, and "Cream of Wheat" is a hot cereal. Below, the cereals are now classified according to one of two classes, "Cold Cereals" or "Hot Cereals."

| Breakfast | Count |
|---|---|
| Cold Cereals | 75 |
| Hot Cereals | 31 |

**Table 11.2. Valid classification scheme**

**12. Make up a data set, consisting of the heights and weights of six children, in which one of the children is an outlier with respect to one of the variables, but not the other. Then alter this data set so that the child is an outlier with respect to both variables.**

In the table below, Child #1 is an outlier with respect to Weight only. All children in the table are close in Height differing at most by 9 inches. However, all children except for Child # 1 are close in Weight differing at most by 7 pounds. Child #1 is an outlier as the Weight differs by 18 pounds from the second-heaviest child (Child #6), making this right-tailed difference in Weight greater than the entire Weight range for the other five children.

| Child | Height (in) | Weight (lbs) |
|---|---|---|
| 1 | 49 | 100 |
| 2 | 50 | 75 |
| 3 | 52 | 77 |
| 4 | 55 | 79 |
| 5 | 57 | 80 |
| 6 | 58 | 82 |

**Table 12.1. Heights & Weights of Children – Weight-only outlier**

In the table below, Child #1 is an outlier with respect to both Height and Weight. All children except for Child #1 in the table are close in Height differing at most by 8 inches and are close in Weight differing at most by 7 pounds. Child #1 is an outlier for both Height and Weight as the Height differs by 14 inches from the second-shortest child (Child#2) (which is greater than the entire Height range of the other five children), and the Weight differs by 18 pounds from the second-heaviest child (Child #6) (which is greater than the entire Weight range of the other five children).

| Child | Height (in) | Weight (lbs) |
|---|---|---|
| 1 | 36 | 100 |
| 2 | 50 | 75 |
| 3 | 52 | 77 |
| 4 | 55 | 79 |
| 5 | 57 | 80 |
| 6 | 58 | 82 |

**Table 12.2. Heights & Weights of Children – Height and Weight outlier**

**Use the following stock price data (in dollars) for Exercises 13–18**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **10** | **7** | **20** | **12** | **75** | **15** | **9** | **18** | **4** | **12** | **8** | **14** |

**Table A. Stock prices**

**13. Calculate the mean, median, and mode stock price.**

The ***mean*** is calculated as the sum of the data points divided by the number of points as follows:

Mean Stock Price = (10+7+20+12+75+15+9+18+4+12+8+14) / 12 = 204 / 12 = $**17**.

The ***median*** is calculated by placing the prices in order and (a) selecting the middle value if the number of points is odd, or (b) taking the average of the two middle values if the number of points is even. Since we have twelve points, median is calculated as follows:

Median Stock Price = mean of center values {4,7,8,9,10,**12,12**,14,15,18,20,75} = 24/2 = $**12**.

The ***mode*** is calculated as the value that occurs the most often in the set and is calculated as follows:

Mode Stock Price = highest frequency of {4,7,8,9,10,**12,12**,14,15,18,20,75} = $**12**.

**14. Compute the standard deviation of the stock price. Interpret what this number means.**

The ***standard deviation*** represents the expected distance of a point chosen at random from a data set to the center of that set and is calculated by taking the square root of the ***variance***. The variance is the average of the sum of squared distances of each point from the data-set mean. Given that the mean is \$17 (see Exercise #13) for this set, the variance for the set of stock prices is calculated as follows:

Stock Price Variance (Var) =

$(4-17)^2+(7-17)^2+(8-17)^2+(9-17)^2+(10-17)^2+(12-17)^2+(12-17)^2+(14-17)^2+(15-17)^2+(18-17)^2+(20-17)^2+(75-17)^2 =$

$(-13)^2 + (-10)^2 + (-9)^2 + (-8)^2 + (-7)^2 + (-5)^2 + (-5)^2 + (-3)^2 + (-2)^2 + (1)^2 + (3)^2 + (58)^2 =$

$169 + 100 + 81 + 64 + 49 + 25 + 25 + 9 + 4 + 1 + 9 + 3364 = 3900 / 12 =$ **325 $\$^2$**.

Taking the square root of the Variance, the Standard Deviation (SD) is calculated as follows:

Stock Price Standard Deviation (SD) of Stock Price = $\sqrt{(325)}$ = **±\$18.03**.

Since the mean is \$17 and the standard deviation is plus/minus \$18.03, the expected price of a stock drawn at random from the set of twelve stocks is expected to lie mathematically between (\$17–\$18.03) = **-\$1.03** (i.e. \$0.01 since we assume that a stock price can never be less than one penny USD) and (\$17+\$18.03) = \$**35.03**.

As we can see, each stock with the exception of the one priced at \$75 is priced within this range.

**15. Find the min-max normalized stock price for the stock worth \$20.**

Min-Max normalization scales an observation relative to the data-set's range resulting in a value between 0 and 1 (this value has no units) and is formulated as follows:

$$\mathbf{MinMaxX_i = [X_i – Min(X)] / [Max(X) – Min(X)]}$$

Therefore, the min-max normalized stock price of \$20 is calculated as follows:

MinMax(\$20) = (\$20 - \$4) / (\$75 - \$4) = (\$16) / (\$71) = **0.2254**.

**16. Calculate the midrange stock price.**

The midrange stock price is the central price for the entire price range and is formulated as follows:

$$\textbf{MidRangeX = [Max(X) + Min(X)] / 2}$$

For the problem at hand we have as follows:

$$\text{MidRangeX} = (\$75 + \$4) / 2 = (\$79) / 2 = \mathbf{\$39.5}$$

**17. Compute the Z-score standardized stock price for the stock worth $20.**

Z-Score standardization scales an observation where the mean value is zero, the SD is 1 and most values lie between -4 and 4 (this value has no units) and is formulated as follows:

$$\textbf{Z-Score(X) = [X}_\textbf{i}\textbf{ – Mean(X)] / |SD(X)|}$$

Given the mean of $17 (see Exercise #13) and |SD| of 18.03 (see Exercise #14), The Z-Score for the stock price of $20 is calculated as follows:

$$\text{Z-Score}(\$20) = (\$20 - \$17) / \$18.03 = (\$3) / \$18.03 = \mathbf{0.1664}.$$

Please note that this value makes sense as it is slightly greater than zero just as $20 is slightly greater than $18.03.

**18. Find the decimal scaling stock price for the stock worth $20.**

Decimal standardization scales an observation to a value between -1 and 1 (this value has no units) and is formulated as follows:

$$\textbf{Decimal(X}_\textbf{i}\textbf{) = X}_\textbf{i}\textbf{ / 10}^\textbf{d}$$

where d is the number of digits in the observation in the data set having the largest absolute value. Since the largest stock price is $75, d = 2 as there are two digits in this price. The decimal standardization is then calculated as follows:

$$\text{Decimal}(\$75) = \$75 / \$10^2 = \$75 / \$100 = \mathbf{0.75}$$

**19. Calculate the skewness of the stock price data.**

Skewness is the lack of normalization of a Z-Score-standardized distribution and is measured using the following formula:

**Skewness = 3 [Mean(X) – Median(X)] / SD(X)**

Given the mean of $17 and median of $12 (see Exercise #13), and an SD of $18.03 (see Exercise #14), the skewness for the stock price distribution is calculated as follows:

Skewness = 3 [$17 - $12] / $18.03 = 3[$5] / $18.03 = $15 / $18.03 = **0.8319**.

We observe that this distribution is right-skewed since a right-skewed distribution has a mean that is greater than its median yielding a positive skewness value. In contrast, a left-skewed distribution will have a mean that is less than its median and thus a negative skewness value.

**20. Explain why data analysts need to normalize their numeric variables.**

Data analysts need to normalize their numeric variables as it places all variables on the same scale. Normalizing all variables to the same scale is critical when performing operations that are sensitive to data variation or ***spread*** so that variables having larger variations do not adversely overpower variables having smaller variations. Most (if not all) analytic operations involving linearization (e.g. Regression, PCA, MANOVA, etc.) are sensitive to data spread.

**21. Describe three characteristics of the standard normal distribution.**

The three main characteristics of the Standard Normal Distribution are as follows:

- The mean is zero
- The SD is 1
- It is symmetric (equal and opposite in shape and size) about the mean and normal (the mean has the highest frequency, and frequency decreases symmetrically as distance from the mean increases).

**22. If a distribution is symmetric, does it follow that it is normal? Give a counterexample.**

If a distribution is symmetric, it is not guaranteed to be normal. In order for a distribution to be normal it has to have a single expected value (i.e. the value with the highest frequency).

A classic counterexample is the Uniform Distribution, which is symmetric about the center of its interval, yet since it all values on the interval occur with equal frequency, it has an infinite number of expected values making it non-normal.

**23. What do we look for in a normal probability plot to indicate non-normality?**

A normal probability plot is simply a plot of the quantiles of a given distribution to the quantiles of the Standard Normal Distribution. If the quantiles are approximately equal, then the plot will approximate a straight line indicating that the given distribution is normal.

In contrast, if the quantiles of the distribution are not equal to the Standard Normal Distribution, then the plot will not approximate a straight line indicating non-normality.

**Use the stock price data for Exercises 24–26.**

**24. Do the following:**

**a. Identify the outlier.**

The outlier is the stock price of \$75. The difference from the next-closest stock price (\$20) is \$55, which is nearly 3.5X larger than the entire range of the other eleven stocks (i.e. \$16).

**b. Verify that this value is an outlier, using the Z-score method.**

We can also verify that \$75 is in fact an outlier using the Z-score method. The Z-score for this stock is calculated using our mean of \$17 (see Exercise #13) and our SD of \$18.03 (see Exercise #14) as follows:

Z-Score(\$75) = (\$75 - \$17) / \$18.03 = (\$58) / \$18.03 = **3.2169**.

Since a Z-score that is less than -3 or greater than 3 is considered an outlier, we conclude that stock price \$75 is an outlier as its Z-score is 3.2169 which is greater than 3.

**c. Verify that this value is an outlier, using the IQR method.**

We can also verify that \$75 is in fact an outlier using the Inter-Quartile Range or IQR method. The quartiles are determined by placing the stock prices in ascending order and dividing them onto four parts as follows:

The ordered stock prices are: {4,7,8,9,10,12,12,14,15,18,20,75}, and since there are an even number of values, we partition as {4,7,8,9,10,12} and {12,14,15,18,20,75}

The quartiles are then determined as follows:
Q1 = {4,7,**8**,9,10,12} = \$8
Q3 = {12,14,**15**,18,20,75}= \$15

We then calculate IQR = Q3 – Q1 as follows:

IQR = $15 – $8 = **$7**

If an observation is an outlier, then it will have a value that is less than Q1 - 1.5IQR or a value greater than Q3 + 1.5IQR. We then calculate the upper and lower boundary values for the stock price set as follows:

LowerBound = Q1 - 1.5IQR = 8 – 1.5(7) = 8 – 10.5 = **-$2.50**
UpperBound = Q3 + 1.5IQR = 15 + 1.5(7) = 15 + 10.5 = $**25.50**

Since $75 is greater than $25.5, we conclude that $75 is an outlier.

**25. Identify all possible stock prices that would be outliers, using:**

**a. The Z-score method.**

The ordered stock prices are: {4,7,8,9,10,12,12,14,**15,18**,20,75} where the mean is $17 lying between the $15 and $18 stock indicated in bold text, and the SD is $18.03. Working from the left, we have as follows:

Z-Score($4) = ($4 - $17) / $18.03 = (-$13) / $18.03 = **-0.7210**.

Z-Score($7) = ($7 - $17) / $18.03 = (-$10) / $18.03 = **-0.5546**.

Z-Score($8) = ($8 - $17) / $18.03 = (-$9) / $18.03 = **-0.4992**.

Z-Score($9) = ($9 - $17) / $18.03 = (-$8) / $18.03 = **-0.4437**.

Z-Score($10) = ($10 - $17) / $18.03 = (-$7) / $18.03 = **-0.3882**.

Z-Score($12) = ($12 - $17) / $18.03 = (-$5) / $18.03 = **-0.2773**.

Z-Score($14) = ($14 - $17) / $18.03 = (-$3) / $18.03 = **-0.1664**.

Z-Score($15) = ($15 - $17) / $18.03 = (-$2) / $18.03 = **-0.1109**.

Z-Score($18) = ($18 - $17) / $18.03 = ($1) / $18.03 = **0.0555**.

Z-Score($20) = ($20 - $17) / $18.03 = ($3) / $18.03 = **0.1664**.

We already know that $75 is an outlier having a Z-score of **3.2169** (see Exercise #24). However, no other outliers were identified using Z-score standardization.

**b. The IQR method.**

The ordered stock prices are: {4,7,8,9,10,12,12,14,**15,18**,20,75} where we have an IQR of **$7**, a Lower Bound of **$2.5**, and an Upper Bound of $**25.50.**

Therefore, stock prices $75 is once again the only outlier as it is greater than the upper bound of **$25.50**.

**26. Investigate how the outlier affects the mean and median by doing the following:**

**a. Find the mean score and the median score, with and without the outlier.**

The mean for the entire set of stock prices is **$17** (see Exercise #13),, and the mean without the $75 outlier is calculated as follows:

$\text{Mean}_{\text{No\_Outlier}}$ = (10+7+20+12+15+9+18+4+12+8+14) / 11 = 129 / 11 = $**11.73**.

The ***median*** is calculated by placing the prices in order and (a) selecting the middle value if the number of points is odd, or (b) taking the average of the two middle values if the number of points is even. Since we have twelve points, median is calculated as follows:

Median Stock Price = mean of center values {4,7,8,9,10,**12,12**,14,15,18,20,75} = 24/2 = $**12**.

$\text{Median}_{\text{No\_Outlier}}$ = mean of center values {4,7,8,9,10,**12,**12,14,15,18,20} = $**12**.

**b. State which measure, the mean or the median, the presence of the outlier affects more, and why.**

It is obvious that the presence of the outlier affects the mean more than the median. It increases the mean by $5, and has no effect on the median.

For this particular data set, the outlier affects the mean more than the median because the mean determines the numerical center of the data set through interpolation and this data is right-skewed having a large right-tailed outlier. In contrast, the median determines the distributive center of the dataset through physical partitioning and the largest value of the lower half of the data is equal to the smallest value of the upper half of this data set.