Study Guide for Quiz 1

- Review lecture notes. They actually include many sample questions!
- Several questions are based on class discussion. You would be able to answer them if you paid full attention in class!
- Reading the textbook is always a plus! But, no question will come out of the topics not covered in lecture notes.
- No question about R programming language or Tableau functions.

Some important keywords

- Mapping different business questions to different data mining tasks or other data technologies
- Data mining vs. other (conventional) data analytics tools (e.g. data warehouse, OLAP, SQL, etc.)
- Supervised vs. unsupervised data mining
- Understand the differences in several data mining tasks (e.g., classification, regression, …) covered in class
- How do you choose and read charts? Understand when to use each graphic chart (scatter plot, bar chart, histogram, and boxplot) and their characteristics.
    - Which graphic chart should we use to see the relationship between two numeric values?
    - Which graphic chart should we use to see the distribution of people's weight?
- Common data issues in supervised data mining: week 3 content (supervised learning checkpoint keys 1 and 2)
    - Understand the need to have the same phenomenon and the same context of the data
    - Check if we have the right data and it has target variable values
- How to build and interpret a classification tree model?
- ~~Entropy and information gain~~
- ~~Use of model~~

Sample Questions

1) (True/False) Finding customers with repeat purchases within a month is an example of an unsupervised data mining task.

2) A marketing analyst wants to compare monthly sales (in 1000s dollars) figures across five different regions for the past year (12 data points per region). Which type of chart would be most appropriate to visualize to compare the trend?
    a) Pie Chart
    b) Line Chart
    c) Bar Chart
    d) Scatter Plot

3) Which of the following is a common feature of misleading graphs?
    a) Equal spacing of intervals on the axes
    b) A vertical axis starting at zero
    c) Consistent units and labeling
    d) A broken or squished axis that exaggerates differences

4) You are trying to determine whether customers can be grouped based on purchasing behavior without any predefined labels. Which data mining task is most appropriate?
    a) Classification
    b) Clustering
    c) Regression
    d) Profiling

5) Which of the following is a valid reason to keep an outlier in your dataset?
    a) It is due to a known data entry error
    b) It results from a faulty sensor reading
    c) It reflects a rare but correct business event
    d) It violates the assumptions of a statistical model

Answer Key:

1. False
2. (b)
3. (d)
4. (b)
5. (c)