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

A database is a collection of related data from a single source, while a data warehouse is houses data from multiple sources to be analyzed.

1. Define each of the following in terms of Data Mining, give an example of where each item would be used:
   1. Characterization
      1. Summarizes general features of the data and results in a specific ruleset to define a class.
   2. Discrimination
      1. Used to separate between sets of data
   3. Association and Correlation Analysis
      1. Aims to discover associations between data. For example, there’s been a study done that tells us those who buy diapers are also likely to buy beer. These items are associated.
   4. Classification
      1. Categorizes items into a specific class based on pre-defined properties. An example would be classifying shapes based on the number of sides, vertices, etc it has.
   5. Regression
      1. Regression is the measure of a dependent variable based upon another variable. An example of using this could be predicting the height of an individual over time.
   6. Clustering
      1. Clustering observes data and classifies them into groups based on the properties. This is different than classification because there is not a set of rules predefined. An example would be using this method to group similar vehicles together based on their various properties.
   7. Outlier Analysis
      1. Outlier analysis is run to determine if there are any anomalies present in the data. This lets us either increase or decrease our confidence in the data. An example of using this could be monitoring analytics on a website. If there is a sudden surge of traffic, it could be an outlier due to bot traffic or web crawlers.
2. Present an example where data mining is crucial to the success of a business. What data mining functionalities does this business need? Can these needs be completed with simple data queries or statistical analysis?

An example where data mining is crucial for a business is an e-commerce website. These sites analyze lots and lots of data. They’re often looking for patterns between products, so the website can introduce efficient marketing strategies and increase revenue. This can be done using association analysis. In addition, they’re using outlier analysis to track website traffic. This would require more than simple data queries or analysis.

1. What is the difference between data mining and statistics?

Statistics is simply quantifying data and producing attributes about the data. Data mining, on the other hand, is using various techniques to discover trends and pattern within the data.

1. What are the 4 main types of data? Describe an example of each.

Nominal (gender), ordinal (t-shirt size), discrete (number of students in a classroom), continuous (the real numbers).

1. Determine the type of data for each of the following:
   1. Height
      1. Continuous
   2. Gender
      1. Nominal
   3. City of residence
      1. Nominal
   4. Number of children
      1. Discrete
   5. Age
      1. Discrete
   6. Highest level of education completed
      1. Ordinal
   7. Ethnicity
      1. Nominal