* Q1.2 – How is a *data warehouse* different from a *database*? How are they similar?

A data warehouse pulls information from numerous different sources. Prior to utilizing the information, it must be cleaned, integrated, transformed, and loaded under a single schema. This information is usually historical in nature and summarized or aggregated into higher levels than a database. As a result, this data is modeled as a data cube and used for online analytical processing (OLAP). This allows the user to drill-down or drill-up through the data, depending on the what the application or investigation calls for. Therefore, they are usually more complex and involved in more intensive queries than a database, which relies on single transactions via read-write operations on online transactional processing (OLTP). However, a database usually contains more detailed information than a warehouse. Despite these differences, there are some similarities between a data warehouse and database. Both of these systems utilize can SQL to retrieve information. Furthermore, this information is typically stored in a similar structure of rows, columns, indices.

* Q1.3 – Define each of the following *data mining functionalities*: characterization, discrimination, association and correlation analysis, classification, regression, clustering, and outlier analysis. Give examples of each data mining functionality, using a real-life database you are familiar with.
  + **Characterization** – Summarizing data into a target class based on broad descriptions of the data. Typically, this information is displayed in a pie chart, bar chart, etc.
    - *Example – Characterizing products at my company that sold more than $50,000 last year.*
  + **Discrimination** – Comparing features of one target class against one more target class of contrasting characteristics.
    - *Example – Comparing products characteristics at my company that show increasing/decreasing sales trends in the Midwest region against the Western region of the United States.*
  + **Association and Correlation Analysis** – Reviewing frequent patterns in data that have a designated confidence and support of occurring together.
    - *Example – Reviewing sequential patterns of price increases against decreases in orders/demand.*
  + **Classification** – Defining a function/model that describes or defines a not yet labeled target class. Typically, classification creates discrete labels.
    - *Example – Defining suppliers based on product supplied, quantity/capacity, price, size, etc.*
  + **Regression** – Define models of continuous-valued functions and predict missing values.
    - *Example – Customer demand in upcoming days/months/years based on previously collected data.*
  + **Clustering** – Grouping data (not yet labeled and/or classified) into groups to maximize intraclass similarity and minimize interclass similarity.
    - *Example – Group products based on sales/cost/market/etc. to find homogenous groups.*
  + **Outlier Analysis** – Reviewing data for points that are outside of the normally expected values.
    - *Example – Reviewing statistical process control (SPC) data for out of control data points that are outside the normally expected values.*
* Q1.5 – Explain the difference and similarity between discrimination and classification, between characterization and clustering, and between classification and regression.
  + **Discrimination and Classification**:
    - *Similarities* – Both functionalities are descriptive in nature, compare/contrast data, and typically focus on discrete/categorical labels.
    - *Differences* – Discrimination compares/contrasts user-selected classes to describe the differences between classes; however, classification creates a model to assign the data into a specific class based on the characteristics of the given information.
  + **Characterization and Clustering:**
    - *Similarities* – Both are descriptive in nature and group similar bits of information together based on characteristics of the data.
    - *Differences* - Clustering groups data into new, distinct classes/categories that may not have been apparent to the user at first; however, characterization groups data based on user-defined characteristics/criteria.
  + **Classification and Regression:**
    - *Similarities* – Both functionalities define a model/function based on the given data. In addition, a relevance analysis is typically needed prior to creating the model to determine which attributes to include. Both can be modeled as a formula.
    - *Differences* – Classification is descriptive in nature and utilizes discrete/categorical labels. Typically, classification models can be shown as decision trees. However, regression is typically predictive in nature and outputs continuous-value information. Usually, regression can only be modeled as a formula.
* Q1.8 – Describe three challenges to data mining regarding *data mining methodology* and *user interaction issues*.

One challenge faced in data mining is the information and data that is provided. Typically, the data sets are missing values, inconsistent, improperly formatted, and difficult to access. Without proper care and attention to detail in getting these items addressed, the user may generate models that do not truly represent the data. As a result, the conclusions drafted from these analyses will not drive the expected improvements. That is why it is vital for the user to properly clean and pre-process the data to address outliers and remove uncertainty.

Another challenge that may be encountered in data mining is the overall experience. The experience should be interactive and user-friendly. The user should be able to investigate, drill-down, drill-up, slice, and view the data from numerous angles. It should be an exploratory journey to find where items need to be address. Business experts need to drive actionable results from these analyses, not get further confused and uncertain by the information that is provided to them.

Background information is another vital challenge that needs to be overcome when working with data. One user, whom is not familiar with the data, may overlook an insightful pattern that someone with subject matter expertise would find very interesting. Furthermore, a subject matter expert would also be able to explain what assumptions and/or restrictions are realistic and need to be accounted for. Therefore, it is vital to gain the appropriate knowledge ahead of time or ensure the proper team members are involved when driving a project forward.

* Q1.9 – What are the major challenges of mining a huge amount of data (e.g. billions of tuples) in comparison with mining a small amount of data (e.g. data set of a few hundred tuple)?

There are several major challenges that arise as the amount of data is scaled up for mining. First, there is the challenge of meeting the timing required of the specific application. The algorithm/model must not get bogged down with the additional data load. This could cause critical failures if information is required at a specific time-interval. Secondly, as the complexity of the algorithms increase, the required computing power also increases. This can be addressed by utilizing parallel algorithms, cloud computing, cluster computing, or incremental data mining to ease the workload and decrease costs. As the data scales up, there is also the likelihood that the diversity of data types/locations will increase. This means that text data, video data, global depositories, and numerous other sources must be mined and pre-processed appropriately before analyzing. Therefore, the model must be robust enough to handle all of these different data types, while meeting all other criteria. In regard to social challenges, there is also the issue of data privacy as you scale up. A user does not want to infringe upon someone’s privacy and take their personal data, so this much also be strictly monitored when collected larger and larger amounts of data. With these in mind, a user must ensure their model is efficient, scalable, performance-focused, and optimized as the amount of data increases.

