1. 1.1 What is data mining?

   (a) Is it another hype?

   Nope. Many people treat data mining knowledge discovery from data (KDD), while others view data mining an essential step in this KDD process. KDD process is an iterative sequence of data cleaning, data integration, data selection, data transformation, data mining (intelligent methods to extract data patterns), pattern evaluation, and knowledge presentation.

   (b) Is it a simple transformation or application of technology developed from databases, statistics, machine learning, and pattern recognition?

   Nope. Data mining has incorporated these technology, and it is developing towards an interdisciplinary field.

   (c) Data mining is the result of the evolution of database technology. DM is also the result of the evolution of machine learning research? based on the historical progress? Address the same for the fields of statistics and pattern recognition.

      i. Based on historical progress of machine learning, DM is also its evolution result.
      ii. DM is the evolution result of statistics and pattern recognition. From the historical prospective,

   (d) Describe the steps involved in data mining viewed as KDD.

   KDD process is an iterative sequence of data cleaning(remove noise, NAs and inconsistent data), data integration(combine multiple data sources), data selection(data relevant to the analysis task are retrieved from the database), data transformation(summary or aggregation, data are transformed and consolidated into forms appropriate for mining), data mining (intelligent methods to extract data patterns), pattern evaluation(perform interestingness measurements to identify the truly interesting patterns) , and knowledge presentation(visualize or present mined knowledge to users).

2. 1.2 Difference between data warehouse and database. Similarities?

   (a) Differences: data warehouse integrates data originating from multiple sources and various timeframes then consolidates data in multidimensional materialized data cubes; while database is a collection of interrelated data, and a relational database is a collection of tables, including attributes and tuples. Relational database is kind of composed with 2-dimensional tables.

   (b) Similarities: Both are structured large collections of data.

3. 1.4 Present an example where DM is crucial to the success of a business. What DM functionalities does this business need (e.g., think of the kinds of pattern that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis?

   (a) Example: Data mining can find potential client. For example, the recommendation from amazon according to people's personalized habits of consuming.

(b) Functionalities used: business need characterization and discrimination pattern to get information, for example, about general profiles about customs and categorize them to make better pinpoint sale plans for them. Association and correlation are also needed by business for the purpose of making 'purchased-together' sale plan. Also, classification can generate the decision tree for business predicting. Cluster Analysis and outlier analysis are also widely used. Almost all DM functionalities can be take advantage of by business planers. AllElectronics in the textbook give excellent examples.

(c) Nope. Data query processing can only allow retrieval of specified subsets of the data. Simple statistical analysis is also not data mining

4. 1.5 Explain the difference and similarity between discrimination and classification, between characterization and clustering, and between classification and regression.

(a) Discrimination and classification: Data **discrimination** is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.(e.g. Comparison general features of software product with sales that increased by 10% last year Against decreased by at least 30% during the same period.) Also, discrimination description should include comparative measures that help to distinguish between the target and contrasting classes. **Classification** is the process of finding a model(or a function) that describes and distinguishes data classes or concepts. The model is used to predict the class label of objects for which the class label is unknown. Therefore, similarity of discrimination and classification is that their purpose of distinguish data is the same; the difference between them is that discrimination is functioned as descriptive while classification is functioned as predictive with certain models.

(b) Characterization and clustering: Data characterization is a summarization of the general characteristics or features of a target class of data. (e.g. characterize the software products with sales that increased by 10% in the previous year, the data related to such products can be collected by executing an SQL query on the sales database.) Clustering analyzes data objects without consulting class labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.(e.g. cluster data with respect to customer locations in a city). Therefore, the similarity of characterization and clustering is that both of them tend to summarize data with target purpose and get information from a certain class of data; while the difference is that characterization use class-labeled(training) data sets, clustering analyze data objects without consulting class labels.

(c) Classification and regression: **Classification** is the process of finding a model(or a function) that describes and distinguishes data classes or concepts. The model is used to predict the class label of objects for which the class label is unknown with classification rules, decision trees, mathematical formulae or neural networks etc.. Classification predicts categorical(discrete, unordered) labels. Regression models continuous-valued functions, and it is used to predict missing or unavailable numerical data values rather than (discrete) class labels. Therefore, the similarity of classification and regression is that they are both models for predicting data. The difference is that classification predicts discrete or unordered labels, regression predicts continuous numeric or class label data.

5. 1.7 Outliers. Using fraudulence detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable.

   (a) Method 1: Statistical test that assume a distribution or probability model for the data.

   (b) Method 2: Distance measures where objects that are remote from any other cluster are considered outliers.

   (c) In my opinion, though both methods is quite good to predict outliers, statistical test are more reliable because it can give the distribution while distance measurements from clusters can only tell the very clear outliers, for the data inside the cluster which may be also a fraudulence will not be tested out.

6. 1.9 What are the major challenges of mining billions of tuples in comparison with mining data set of a few hundred tuple?

   Efficiency and scalability. When mining a huge amount of data, data mining algorithms must be efficient and scalable to effectively extract information. The algorithms that can execute in real time is useful. When mining a small amount of data, algorithms might be not make lots of difference with each other since they may take approximately equal time for small data set. While for large amount of data, algorithms and methods to deal with certain data for a targeted purpose would distinguish results.

- UCI Iris DATA interpretation (Using python)

  – I generated table.py to calculate the minimum, maximum, mean, standard error, variation for total 150 tuples, 50 setosa tuples, 50 versicolor tuples, and 50 virginia tuples separately. The result is iris.table, which can also seen from below:

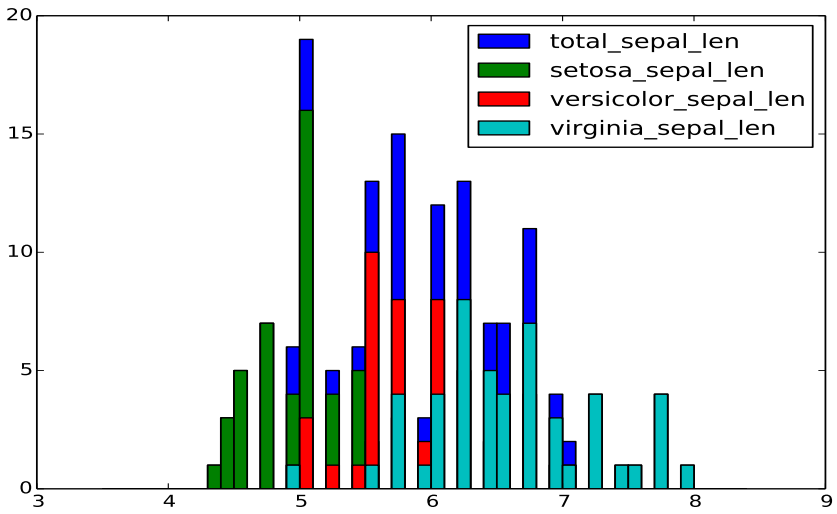
NAN	sepal_len	sepal_wid	petal_len	petal_wid
Total_min	4.3	2.0	1.0	0.1
Total_max	7.9	4.4	6.9	2.5
Total_mean	5.8433333	3.054	3.7586666	1.1986666
Total_sd	0.8253012	0.4321465	1.7585291	0.7606126
Total_var	0.6811222	0.1867506	3.0924248	0.5785315
Setosa_min	4.3	2.3	1.0	0.1
Setosa_max	5.8	4.4	1.9	0.6
Setosa_mean	5.006	3.418	1.464	0.244
Setosa_sd	0.3489469	0.3771949	0.1717672	0.1061319
Setosa_var	0.121764	0.142276	0.029504	0.011264
Versicolor_min	4.9	2.0	3.0	1.0
Versicolor_max	7.0	3.4	5.1	1.8
Versicolor_mean	5.936	2.77	4.26	1.326
Versicolor_sd	0.5109833	0.3106444	0.4651881	0.1957651
Versicolor_var	0.261104	0.0965	0.2164	0.038324
Virginia_min	4.9	2.2	4.5	1.4
Virginia_max	7.9	3.8	6.9	2.5
Virginia_mean	6.588	2.974	5.552	2.026
Virginia_sd	0.6294886	0.3192553	0.5463478	0.2718896
Virginia_var	0.396256	0.101924	0.298496	0.073924

  – From this table, we can immediately know sepal length range from 4.3 to 7.9 cm and has mean 5.84 cm, varies largely. Also, for the three species (setosa, versicolor, and virginia), their min and max don't change much from 4.3~7.9cm. However, versicolor's mean value is much similar to the total mean, while setosa and virginia deviate from 5.84 about the same absolute value around 0.7cm. Moreover, the variance seems mainly originated from virginia because it has the largest variation among the three.

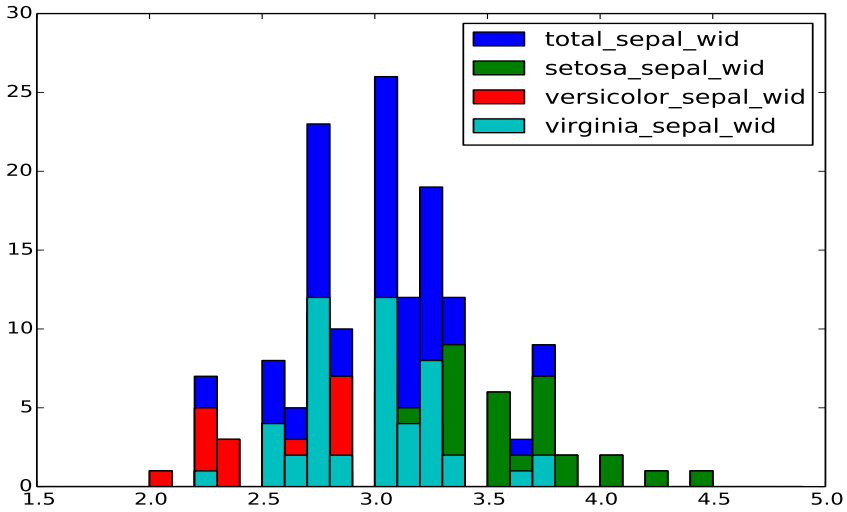
  - (sepal width, petal length and petal width, we can have similar discussion)

- To get better visualization, I plotted total 150 tuples as well as three 50 species' tuples with his.py. By uncomment each block, I got four pdf files which attached in the homework. And we can also see, for sepal length, the histogram has good agreement with the analysis above that virginia varies the most, setosa varies the smallest, mean value should around 5 to 6, range is from 4.3 to 7.9. etc.. The advantage part of histogram is that it can clearly give us the sense of the distribution about the length and width of different species and their individual contribution to the whole graph. The discussions for septal width and petal length and width are similar.

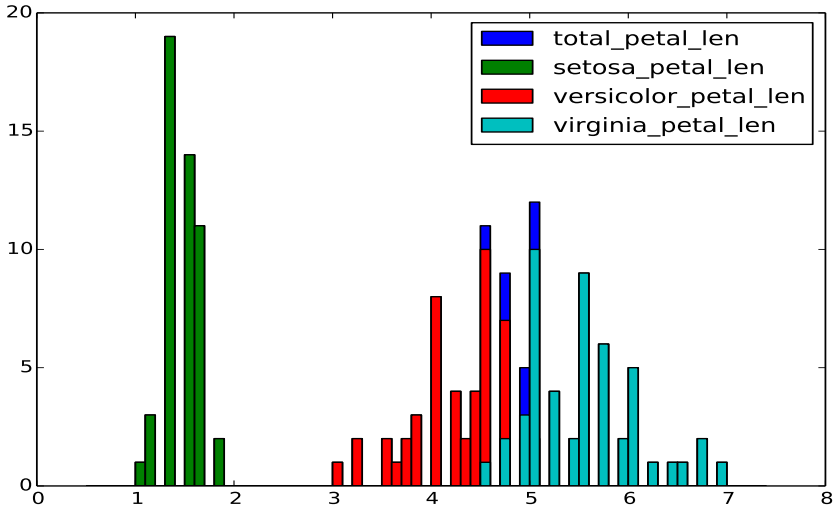
  - Sepal Length



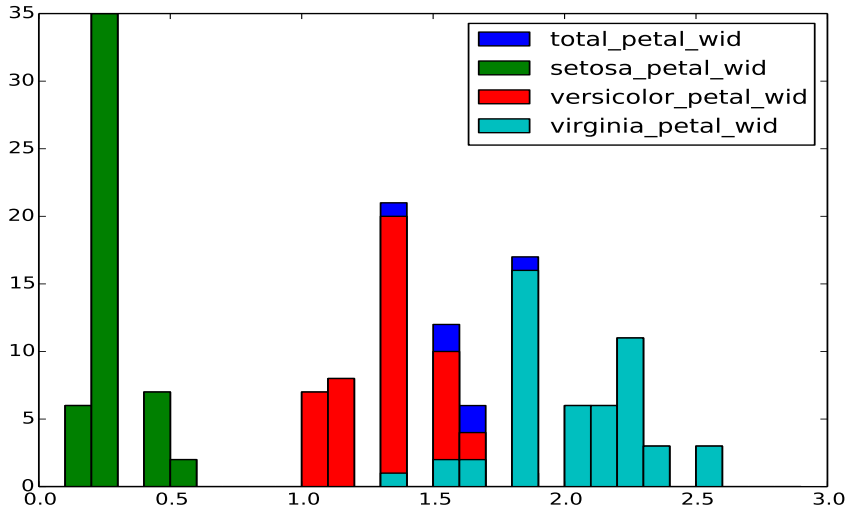
  - Sepal Width



* Petal Length



* Petal Width



  - Except from histograms, numpy can also make 'clustering' by numpy.historgram2d plot. In that plot, for example it can combine width and length for sepal together, also for petal. It will give us more detailed and clear sense about how to distinguish these three 'similar' species – setosa, versicolor, and virginia.

- References:

  - Data Mining: Concepts and Techniques, 3/E Jiawei Han, Micheline, Kamber, and Jian Pei.
  - Numpy Official manual.
  - Stackoverflow.