UECS3213 / UECS3453 DATA MINING

SESSION: January 2019

TUTORIAL 2
Chapter 2a - Data
Chapter 2b - Data Exploration

1. Classify the following attributes as *binary*, *discrete*, or *continuous*. Also classify them as *qualitative* (nominal or ordinal) or *quantitative* (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

   Example: Age in years. Answer: Discrete, quantitative, ratio

   a) Time in terms of "AM" or "PM".
   - Binary, qualitative, ordinal

   b) Brightness as measured by a light meter.
   - Continuous, quantitative, ratio

   c) Brightness as measured by people's judgments.
   - Discrete, qualitative, ordinal (e.g. dark, bright, very bright)

   d) Angles as measured in degrees between 0◦ and 360◦.
   - Continuous, quantitative, ratio

   e) Bronze, Silver, and Gold medals as awarded at the Olympics.
   - Discrete, qualitative, ordinal

   f) Height above sea level.
   - Continuous, quantitative, interval/ratio (depends on whether sea level is regarded as an arbitrary origin)

   g) Number of patients in a hospital.
   - Discrete, quantitative, ratio

   h) ISBN numbers for books. (Look up the format on the Web.)
   - Discrete, qualitative, nominal (ISBN numbers do have order information, though)

   i) Ability to pass light in terms of the following values: opaque, translucent, transparent.

- Discrete, qualitative, ordinal

j) Military rank.
- Discrete, qualitative, ordinal

k) Distance from the center of campus.
- Continuous, quantitative, interval/ratio (depends)

l) Density of a substance in grams per cubic centimeter.
- Continuous, quantitative, ratio

m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)
- Discrete, qualitative, nominal

2. *Proximity* (nearness) is typically defined between a pair of objects. Proximity may be measured in terms of how alike objects are to one another (object similarity) or how unlike they are (object dissimilarity). The term distance measure is often used instead of dissimilarity measure.

a) Define two ways in which you might define the proximity among a group of objects.

Two examples are the following:

i) based on pairwise proximity, i.e., minimum pairwise similarity or maximum pairwise dissimilarity, or

ii) for points in Euclidean space compute a centroid (the mean of all the points) and then compute the sum or average of the distances of the points to the centroid.

b) How might you define the *distance* between two sets of points in Euclidean space?

- One approach is to compute the distance between the centroids of the two sets of points.

c) How might you define the *proximity* between two sets of data objects? (Make no assumption about the data objects, except that a proximity measure is defined between any pair of objects.)

- One approach is to compute the average pairwise proximity of objects in one group of objects with those objects in the other group. Other approaches are to take the minimum or maximum proximity.

3. Suppose that the data for analysis includes the attribute *age*. The age values for the data tuples are (in increasing order) :
13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

a) What is the arithmetic *mean* of the data? What is the *median*?

The (arithmetic) mean of the data is: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 809/27 = 30$. The median (middle value of the ordered set, as the number of values in the set is odd) of the data is: 25.

b) What is the *mode* of the data? Comment on the data's modality (i.e., *bimodal*, trimodal, etc.).

- This data set has two values that occur with the same highest frequency and is, therefore, bimodal. The modes (values occurring with the greatest frequency) of the data are 25 and 35.

c) What is the *midrange* of the data?

- The midrange (average of the largest and smallest values in the data set) of the data is: (70+13)/2 = 41.5

d) Can you find (roughly) the *first quartile (Q1)* (lower quartile) and the *third quartile (Q3)* (upper quartile) of the data?

- The first quartile (corresponding to the 25th percentile) of the data is: 20. The third quartile (corresponding to the 75th percentile) of the data is: 35.
- https://www.varsitytutors.com/hotmath/hotmath_help/topics/interquartile

e) Give the *five-number summary* of the data.

- The five number summary of a distribution consists of the **minimum value, first quartile, median value, third quartile, and maximum value**. It provides a good summary of the shape of the distribution and for this data is: 13, 20, 25, 35, 70.
- Minimum value = 13
- First quartile = 20
- Median = 25
- Third quartile = 35
- Maximum value = 70
- https://www.hackmath.net/en/calculator/five-number-summary

4. Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result:

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

- https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review

a) Calculate the *mean*, *median* and *standard deviation* of age and %fat.

- For the variable age the mean is 46.44, the median is 51, and the standard deviation is 12.85.

- For the variable %fat the mean is 28.78, the median is 30.7, and the standard deviation is 8.99.

b) Draw the *boxplots* for age and %fat.

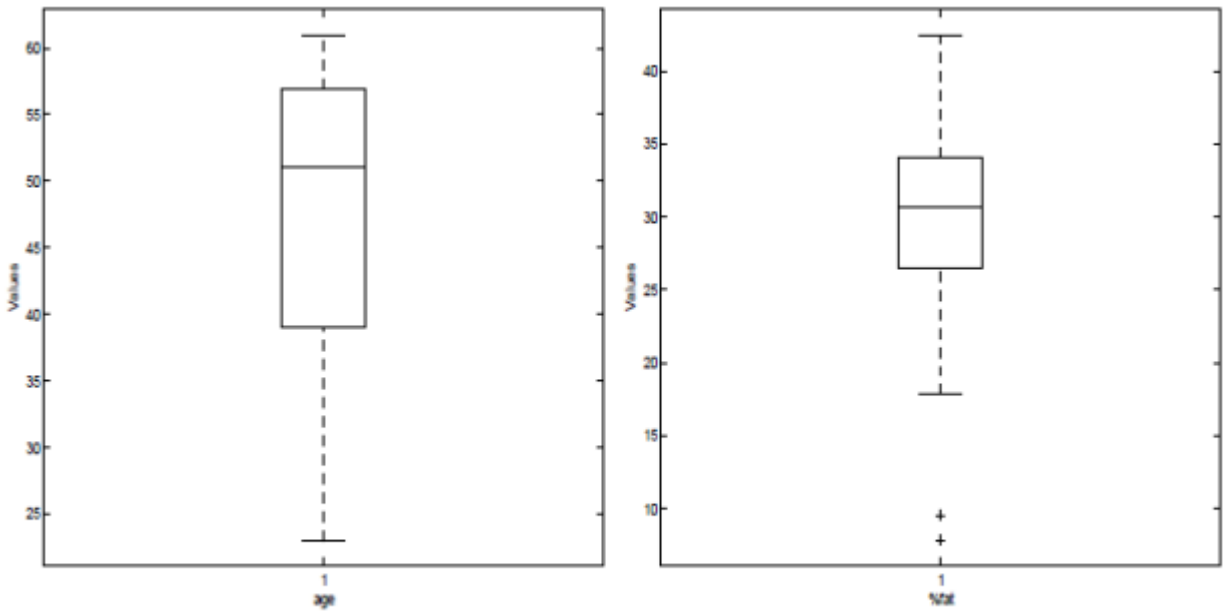


Figure 2.2: A boxplot of the variables *age* and *%fat*

https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review

c) Draw a *scatter plot* and a quantile-quantile / *q-q plot* based on these two variables.

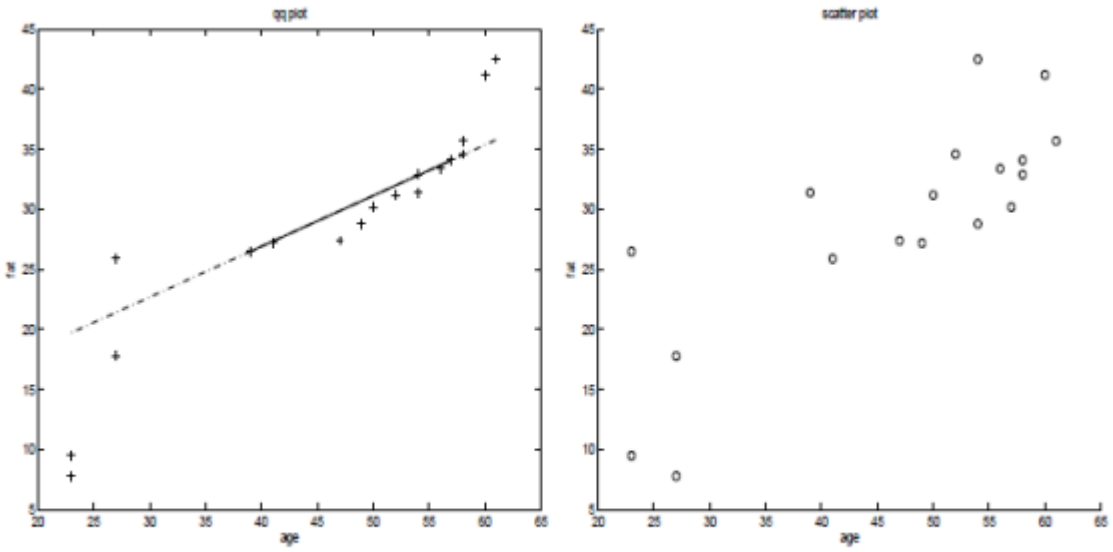


Figure 2.3: A *q-q plot* and a *scatter plot* of the variables *age* and *%fat*

- q-q plot:
- https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot
- https://www.statisticshowto.datasciencecentral.com/q-q-plots/

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. Can be plotted using Excel or other software tool.

d) Calculate the *Pearson correlation coefficient*. Are these two variables positively or negatively correlated?

The Pearson correlation coefficient is 0.82, the variables are positively correlated.
https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

5. Construct a *data cube* from Table 3.1 below. Is this a *dense* or *sparse* data cube? If it is sparse, identify the cells that are empty.

Table 3.1. Fact table

Product ID	Location ID	Number Sold
1	1	10
1	3	6
2	1	5
2	2	22

The data cube is shown in Table 3.2. It is a dense cube; only two cells are empty.

Table 3.1. Fact table

Product ID	Location ID	Number Sold
1	1	10
1	3	6
2	1	5
2	2	22

Table 3.2. Data cube

Product ID \ Location ID	1	2	3	Total
1	10	0	6	16
2	5	22	0	27
Total	15	22	6	43

6. This exercise compares and contrasts some *similarity* and *distance measures*.

a) For binary data, the L1 distance corresponds to the *Hamming distance*; that is, the number of bits that are different between two binary vectors. The *Jaccard similarity* is a measure of the similarity between two binary vectors.

Compute the *Hamming distance* and the *Jaccard similarity* between the following two binary vectors.

- x = 0101010001
- y = 0100011000
- Hamming distance = number of different bits = 3
- Jaccard Similarity = number of 1-1 matches / (number of bits - number 0-0 matches) = 2 / (10 - 5) = 2 / 5 = 0.4
- https://en.wikipedia.org/wiki/Hamming_distance

b) Which approach, Jaccard or Hamming distance, is more similar to the *Simple Matching Coefficient (SMC)*, and which approach is more similar to the *cosine measure*? Explain. (Note: The Hamming measure is a distance, while the other three measures are similarities, but don't let this confuse you.)

- The Hamming distance is similar to the SMC. In fact, SMC = Hamming distance / number of bits.
- The Jaccard measure is similar to the cosine measure because both ignore 0-0 matches.

c) Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)

- Jaccard is more appropriate for comparing the genetic makeup of two organisms; since we want to see how many genes these two organisms share.
- Jaccard Similarity = number of 1-1 matches / (number of bits - number 0-0 matches)

d) If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note that two human beings share > 99.9% of the same genes.)

- Two human beings share >99.9% of the same genes. If we want to compare the genetic makeup of two human beings, we should focus on their differences. Thus, the Hamming distance is more appropriate in this situation.

7. For the following vectors, x and y, calculate the indicated similarity or distance measures.

Cosine similarity: https://en.wikipedia.org/wiki/Cosine_similarity

a) x = (1, 1, 1, 1), y = (2, 2, 2, 2) cosine, correlation, Euclidean

- cos(x, y) = 1, corr(x, y) = 0/0 (undefined), Euclidean(x, y) = 2

b) x = (0, 1, 0, 1), y = (1, 0, 1, 0) cosine, correlation, Euclidean, Jaccard

- cos(x, y) = 0, corr(x, y) = −1, Euclidean(x, y) = 2, Jaccard(x, y) = 0

c) x = (0,−1, 0, 1), y = (1, 0,−1, 0) cosine, correlation, Euclidean

- cos(x, y) = 0, corr(x, y)=0, Euclidean(x, y) = 2

d) x = (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1) cosine, correlation, Jaccard

- cos(x, y) = 0.75, corr(x, y) = 0.25, Jaccard(x, y) = 0.6

e) x = (2,−1, 0, 2, 0,−3), y = (−1, 1,−1, 0, 0,−1) cosine, correlation

- cos(x, y) = 0, corr(x, y) = 0

Discussion Questions

1. *Proximity* is typically defined between a pair of objects.

   a) Define two ways in which you might define the proximity among a group of objects.

- Two examples are the following: (i) based on pairwise proximity, i.e., minimum pairwise similarity or maximum pairwise dissimilarity, or (ii) for points in Euclidean space compute a centroid (the mean of all the points) and then compute the sum or average of the distances of the points to the centroid.

   b) How might you define the distance between two sets of points in Euclidean space?

- One approach is to compute the distance between the centroids of the two sets of points.

   c) How might you define the proximity between two sets of data objects?

- (Make no assumption about the data objects, except that a proximity measure is defined between any pair of objects.) One approach is to compute the average pairwise proximity of objects in one group of objects with those objects in the other group. Other approaches are to take the minimum or maximum proximity.

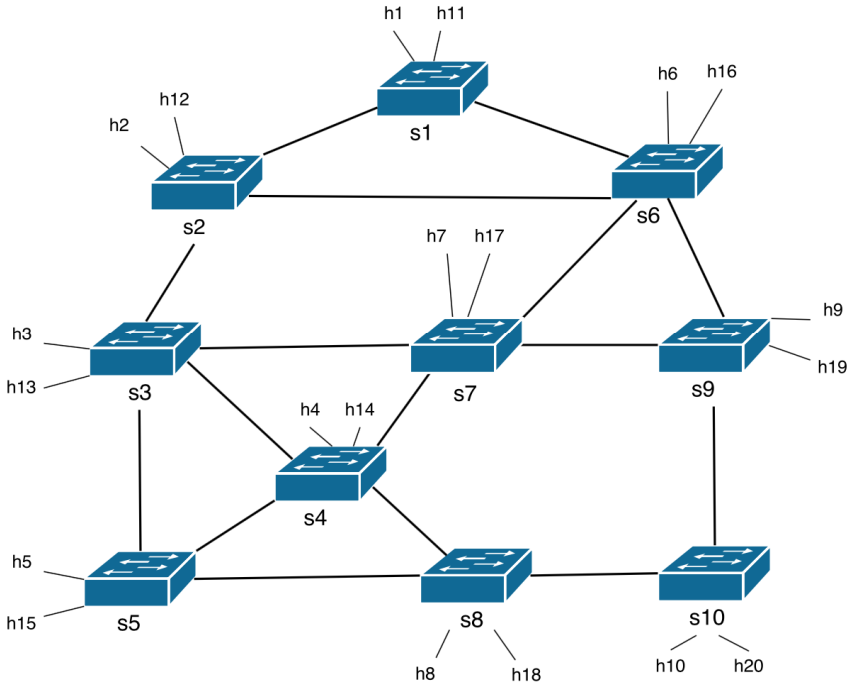
2. Describe how you would create *visualizations* to display information that describes the following types of systems.

   Be sure to address the following issues:

   - **Representation**. How will you map objects, attributes, and relationships to visual elements?
   - **Arrangement**. Are there any special considerations that need to be taken into account with respect to how visual elements are displayed? Specific examples might be the choice of viewpoint, the use of transparency, or the separation of certain groups of objects.
   - **Selection**. How will you handle a large number of attributes and data objects?

   a) Computer networks. Be sure to include both the static aspects of the network, such as *connectivity*, and the dynamic aspects, such as *traffic*.

- The connectivity of the network would best be represented as a **graph**, with the nodes being routers, gateways, or other communications devices and the links representing the connections. The bandwidth of the connection could be represented by the width of the links. Color could be used to show the percent usage of the links and nodes. (light: low percent, dark: high percent)
- Example:

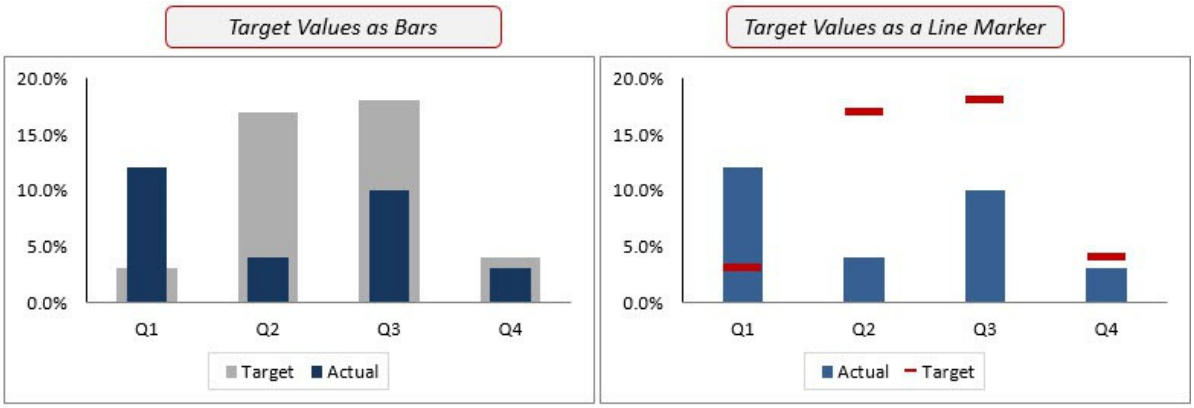


b) The distribution of specific plant and animal species around the world for a specific moment in time.

- The simplest approach is to display each species on a separate **map** of the world and to shade the regions of the world where the species occurs.
- If several species are to be shown at once, then icons for each species can be placed on a map of the world.
- Example:

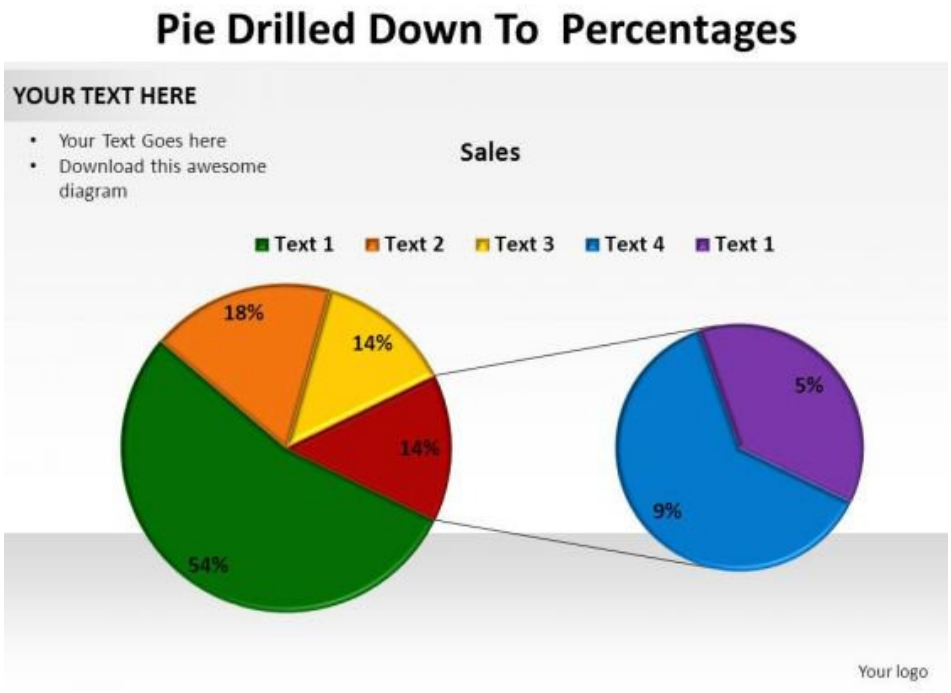
c) The use of computer resources, such as processor time, main memory, and disk, for a set of benchmark database programs.

- The resource usage of each program could be displayed as a **bar plot** of the three quantities.
- Since the three quantities would have different scales, a proper scaling of the resources would be necessary for this to work well. For example, resource usage could be displayed as a percentage of the total.
- Alternatively, we could use three bar plots, one for type of resource usage. On each of these plots there would be a bar whose height represents the usage of the corresponding program. This approach would not require any scaling. Yet another option would be to display a line plot of each program's resource usage.
- For each program, a line would be constructed by (1) considering processor time, main memory, and disk as different x locations, (2) letting the percentage resource usage of a particular program for the three quantities be the y values associated with the x values, and then (3) drawing a line to connect these three points. Note that an ordering of the three quantities needs to be specified, but is arbitrary. For this approach, the resource usage of all programs could be displayed on the same plot.
- Example:



d) The change in occupation of workers in a particular country over the last thirty years. Assume that you have yearly information about each person that also includes gender and level of education.

- For each gender, the occupation breakdown could be displayed as an array of pie charts, where each row of pie charts indicates a particular level of education and each column indicates a particular year. For convenience, the time gap between each column could be 5 or ten years.
- Alternatively, we could order the occupations and then, for each gender, compute the cumulative percent employment for each occupation. If this quantity is plotted for each gender, then the area between two successive lines shows the percentage of employment for this occupation. If a color is associated with each occupation, then the area between each set of lines can also be colored with the color associated with each occupation. A similar way to show the same information would be to use a sequence of stacked bar graphs.
- Example:



3. Explain why computing the *proximity* (distance) between two attributes is often simpler than computing the *similarity* between two objects.

- In general, an object can be a record whose fields (attributes) are of different types. To compute the overall similarity of two objects in this case, we need to decide how to compute the similarity for each attribute and then combine these similarities. This can be done straightforwardly but is still somewhat ad hoc, at least compared to proximity measures such as the Euclidean distance or correlation, which are mathematically well-founded.
- In contrast, the values of an attribute are all of the same type, and thus, if another attribute is of the same type, then the computation of similarity is conceptually and computationally straightforward.

4. Discuss the differences between dimensionality reduction based on aggregation and dimensionality reduction based on techniques such as principal component analysis (PCA) and singular value decomposition (SVD).

- The dimensionality of PCA or SVD can be viewed as a projection of the data onto a reduced set of dimensions. In aggregation, groups of dimensions are combined. In some cases, as when days are aggregated into months or the sales of a product are aggregated by store location, the aggregation can be viewed as a change of scale. In contrast, the dimensionality reduction provided by PCA and SVD do not have such an interpretation.
- https://bigdata-madesimple.com/decoding-dimensionality-reduction-pca-and-svd/

The End