Name: Akond Rahman
Unity ID: aarahman
Subject: CSC 522-HW1

Answer to Q1 (a):
Data mining is the process of analyzing large volumes of data using artificial intelligence techniques, neural networks, and advanced statistical tools to identify meaningful trends, and relationships that can benefit business vendors and the academic community.

Answer to Q1 (b):
The process of data mining encompasses data pre-processing, data visualization, data analysis, real time trend prediction using parallel algorithms, statistical methods, machine learning methods, artificial intelligence techniques etc. On the other hand, statistical analysis works on structured datasets, includes no pre-processing, and does not use methods and techniques from other disciplines such as artificial intelligence, and machine learning.

Answer to Q1 (c):
1. Size of data, e.g. large scale datasets with sizes of gigabytes, or petabytes
2. Multiple dimensionality, e.g. colored satellite images
3. Heterogeneous properties within a single dataset, e.g. collection of webpages with structured, and semis-structured links, text, media content etc.
4. Location of data, e.g. datasets such as Github usage data is stored in a remote server which needs to be processed using specific APIs, no longer possible with traditional techniques.
5. Non traditional analysis, e.g. with the evolution of web and social media, understanding social relationships from social networking sites necessitates to perform graph mining on a large scale of data.

Answer to Q2 (a):
1. A climate dataset which consists temperature measurements of various locations
2. A satellite image dataset containing images of a certain location.
3. NYC Taxi Trip dataset that can be used to investigate which routes of NYC is traversed most at what times of day
4. Legal documents dataset, a dataset maintained by a law firm, that contains legal documents of institutions from different domains such as education, business, and IT
5. Software developer dataset, e.g. the Github dataset that collects all the software projects, commits, issues, pull requests etc. of all the software developers that are using Github

Answer to Q2 (b):

1. Climate dataset: For a complex dataset like this a combination of techniques is needed. For example, support vector machines for classification, clustering for grouping different regions based on temperature, pressure etc., and outlier detection using change detection. To classify regions based on temperature we can use artificial neural networks because artificial neural networks can learn from different layers of machine learning algorithms and perform classification.
2. Satellite image dataset: Depends on the task. For example, for classifying regions we can use different variants of clustering.
3. NYC Taxi Trip dataset: Depends on task. To predict rise of taxi price, the best way would be find out the most correlating factors to price such as time of the year, location, hours of the day using Pearson or Spearman correlation, and then predicting the rise of price from these factors using decision trees.
4. Legal documents dataset: Combination of natural language processing algorithms, and sequence mining to observe interesting patterns that associate over a selected number of topics.
5. Software developer dataset: Depends on task. For example, to study relationships between collaborators of a project, and how these collaborators collaborate with other members of a different project graph mining techniques can be helpful for large dataset.

Answer to Q3 (a):

1. Nominal – distinctness
2. Ordinal - order
3. Interval – differences between values
4. Ratio - differences and ratios

Answer to Q3 (b):

1. Nominal – permutation of values e.g. re-arranging employee IDs
2. Ordinal - transformations that preserve the order, e.g. 'high', and 'low' can eb expressed as a Boolean one and zero.
3. Interval – an equation that converts one value to another, for example converting temperate readings from Celsius to Fahrenheit.
4. Ratio - a ratio-based equation such as newValue=C*oldValue, where C is a constant

Answer to Q3 (c):
An asymmetric attribute is an attribute for which non-zero values are only considered meaningful. For example, in a student course management dataset, for a certain student only the courses that he/she has taken matters.

Answer to Q4 (a):
Original: {55, 23, 28, 32, 18, 68, 72, 89, 98, 100}
Transformed, not rounded to integers: { 115.06098, 15.54878, 31.09756, 43.53659, 0.00000, 155.48780, 167.92683, 220.79268, 248.78049, 255.00000 }

Transformed, rounded to integers : { 115, 16, 31, 44,   0, 155, 168, 221, 249, 255}

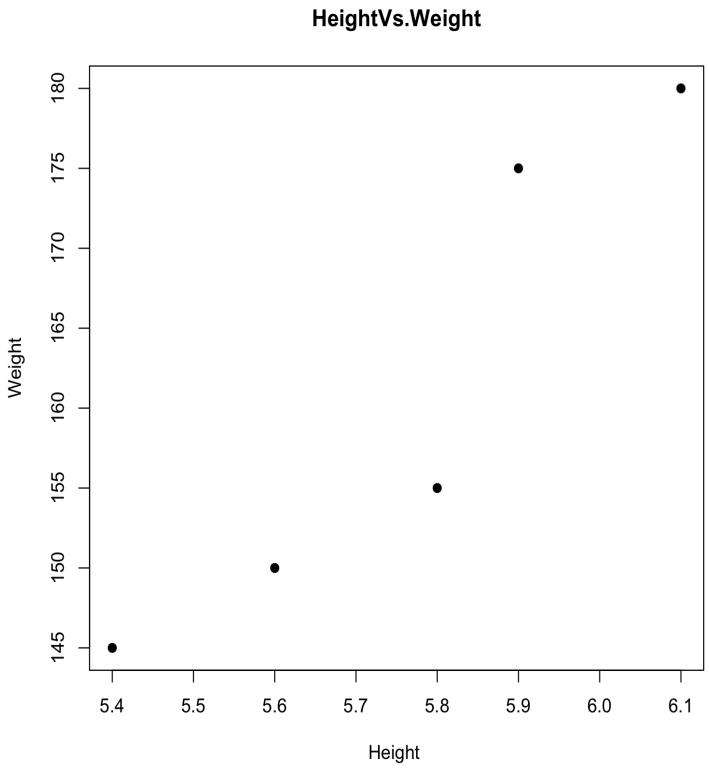
Answer to Q4 (b):
Original: {55, 23, 28, 32, 18, 68, 72, 89, 98, 100}
Transformed, not rounded to integers: {-0.1041811, -1.1144220, -0.9565719, -0.8302918, -1.2722722, 0.3062293, 0.4325094, 0.9691999, 1.2533301, 1.3164702}
Transformed: {0, -1, -1, -1, -1, 0, 0, 1, 1, 1}

Answer to Q5 (a):

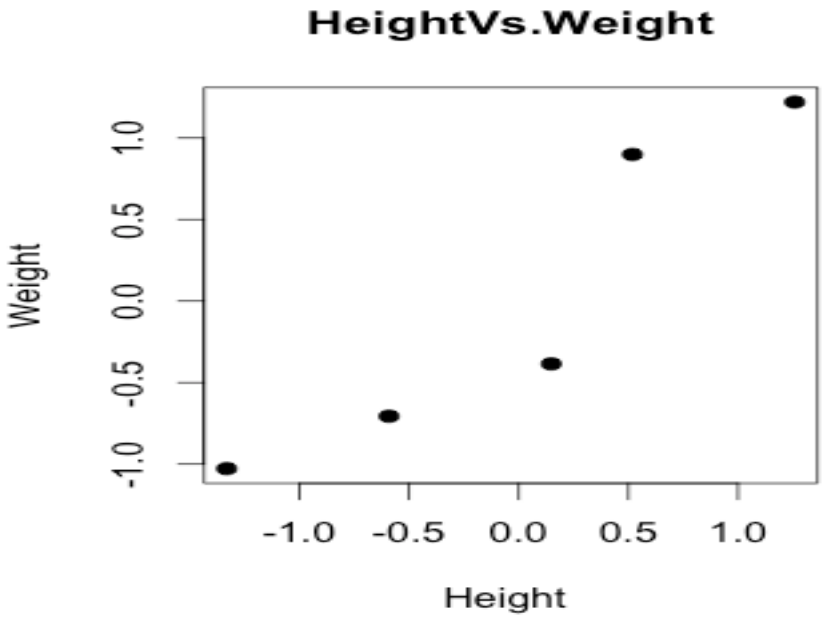
1. Climate dataset: Weighted similarity metric based on distance
2. Satellite image dataset: Distance based similarity metric e.g. Mahanabolis
3. NYC Taxi Trip dataset: Correlation based weighted similarity metric
4. Legal documents dataset: Cosine similarity to find similar documents
5. Software developer dataset: Correlation analysis

Answer to Q5 (b):
1.

Without normalization



Applied Z-Score normalization



2.

Using Euclidian distance, without applying Z-score normalization

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.000000	5.003998	5.003998	30.004166	25.001800
[2,]	5.003998	0.000000	10.007997	25.001800	20.000250
[3,]	5.003998	10.007997	0.000000	35.006999	30.004166
[4,]	30.004166	25.001800	35.006999	0.000000	5.003998
[5,]	25.001800	20.000250	30.004166	5.003998	0.000000

Using Euclidian distance, applied Z-score normalization.

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.000000	0.8068693	0.8068693	2.6713288	1.9519724
[2,]	0.8068693	0.000000	1.6137385	1.9519724	1.3365893
[3,]	0.8068693	1.6137385	0.000000	3.4298506	2.6713288
[4,]	2.6713288	1.9519724	3.4298506	0.000000	0.8068693
[5,]	1.9519724	1.3365893	2.6713288	0.8068693	0.000000