CSC591: Foundations of Data Science
HW1: Exploratory Analysis, Basic Probability, Random Variables and Probability Distributions

Q1. Describe concisely various forms of data preparation (or preprocessing). (10 points)

Data cleaning, data transformation, feature selection, dimensionality reduction, missing value imputation, noise identification, etc. You should define and explain more on each category.

Data Transformation: Data transformations are general preprocessing techniques used to make data more efficient for subsequent data mining. Examples include: smoothing, aggregation, normalization, discretization, etc.

Feature Selection/Dimensionality Reduction: Datasets sometimes have redundant or uninformative features, which can slow down and/or reduce the accuracy of data mining algorithms. Feature selection consists of detecting and removing these features, and dimensionality reduction combines original features into a smaller number of more useful features.

Noise elimination or outlier detection: remove the noise data to increase the accuracy of the analysis, or detect outliers to determine whether the outliers are important or meaningless.

Q2. **A company wants to know the customer satisfaction and conducts survey over 100 customers. The survey form includes the following questions. (10 points)**

Survey Form:

a) Are you: Male Female (Circle one of the choice)
b) How old are you? _____ (in years)
c) How much do you spend on groceries? ____ (in $$$$.$$)
d) How much do you spend on soft drinks? ___ (in $$$$.$$)
e) Which soft beverage do you prefer? ____ (Coke, Pepsi, Dr. Pepper, …)
f) How satisfied are you with diet beverages? ____ (Very satisfied, Satisfied, Not Satisfied)
g) How likely are you to buy 6-pack diet coke?___ (Very likely, Likely, Not Likely, Very unlikely)

Assume that all surveyed customers returned survey forms with correct answers.

Answer the following questions:

Your objective is to:

1. Design the database (one table) and enter the data. Show the table with few sample data entries.

Gender	Age	Groceries Spending	Drinks Spending	Drink Preferrred	Diet Satisfied	Buy diet Coke
Male	25	100	50	Coke	Not Satisfied	Very likely
Male	45	120	20	Pepper	Very Satisfied	Likely
Female	30	80	30	Pepsi	Satisfied	Likely

2. For each resulting column (attribute), list the type of attribute (in terms of Nominal, Ordinal, Interval, Ratio)
   a. Gender: nominal ({M, F})
   b. Age: ratio
   c. GrocerySpending: ratio
   d. DrinkSpending: ratio
   e. DrinkPreferred: nominal ({Coke, Pepsi, …})
   f. DietSatisfed: ordinal ([NotSat, Sat, VerySat])
   g. BuyDietCoke: ordinal ([VUL, NL, L, VL])

3. For each attribute, what kind of summary (statics) make sense?
   a. Gender: percentage(% male, % female), frequency, mode
   b. Age: mean, median, SD, range
   c. GrocerySpending: same as Age
   d. DrinkSpending: same as Age
   e. DrinkPreferred: same as Gender
   f. DietSatisfed: percentage, frequency, mode
   g. BuyDietCoke: same as DietSatisfed
4. For each attribute, what kind of graphical representation makes most sense? (e.g., pie chart, bar chart, …)
   a. Gender, DrinkPreferred, DietSatisfed, BuyDietCoke: pie chart or bar chart
   b. Age, GrocerySpending, DrinkSpending: histogram or density plot

Q3. (**10** points) Define (precise; one or two sentences) the following:

(a) Statistics: A mathematical way to analyze, interpret, present, and organize data
(b) Population: the total set of actually existing objects or events, or a hypothetical and potential infinite set of objects or events
(c) Sample: a finite set of data collected from a statistical population
(d) Event space: the set of all possible outcomes, equivalent to sample space
(e) Event: any subset of the sample space
(f) Random variable: a numerical quantity of outcomes of an experiment
(g) Experiment: an analysis method for testing different assumptions by trial under constructed and controlled conditions. It has three commons. First, it has more than one outcome; Second, each possible outcome can be specified in advance; Third, outcome depends on chance
(h) Discrete data: data that can only take a finite or countably infinite set of values
(i) Continuous data: data that can take infinitely, uncountable values
(j) Mean: the sum of the numerical values of each and every observations divided by the total number of observations.
(k) Median: the value separating the higher half of a data sample, or a population, or a probability distribution
(l) Variance: the expectation of the squared deviation of a random variable from its mean

Q4. (10 points) The following data shows body temperature readings in degree F of 10 patients. (i) Based on these readings, is a body temperature of 104.2 degree F is unusual? [Hint: Anomaly, Outlier, or Unusual value is defined as Mean ± 2 Standard deviations].

(ii) Based on general medical knowledge, is 104.2 deg F is unusual? Also comment if the definition of anomaly or unusual value defined as above is useful in this analysis?

98.8 98.4 98.2 98.1 99.0 98.9 99.2 98.3 98.2 98.8

Solution:

(i) Mean = 98.59; Standard deviation = 0.3929
Because 104.2 > 98.59 + 2 * 0.3929 = 99.3757, so it is unusual
(ii) Based on general medical knowledge, 104.2 is unusual. The above definition of anomaly is useful in the analysis, as it really detects unusual temperature as anomaly.

(iii) Using z score formula, we can get z score is

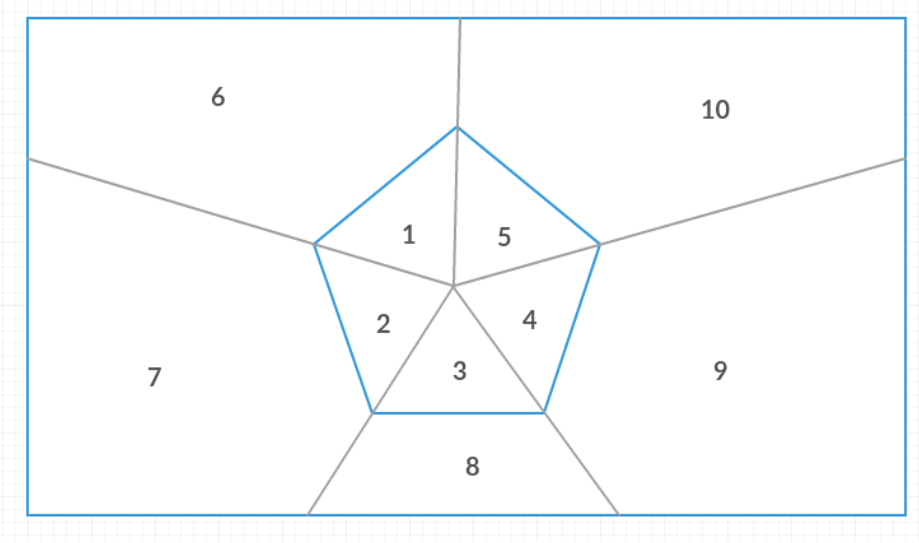
$$z = \frac{x - \mu}{\sigma} = \frac{104.2 - 98.59}{0.3929} = 14.2784$$

Q5. (10 points) Law of total probability: Suppose $C_1, C_2, \ldots, C_m$ are disjoint events such

that $C_1 \cup C_2 \cup \ldots \cup C_m = \Omega$. The probability of an arbitrary event A can be expressed as: $P(A) = P(A|C_1)P(C_1) + P(A|C_2)P(C_2) + \ldots + P(A|C_m)P(C_m)$.

Solution:

Illustrate this law using Venn diagram (for m=5) and derive P(A) using this Venn diagram



For this Venn diagram:
The events are represent as combination of colors:
A: {1,2,3,4,5}
C_1:{1,6}
C_2:{2,7}
C_3:{3,8}
C_4:{4,9}
C_5:{5,10}

Derivation:

$$P(A|C_1)P(C_1) + P(A|C_2)P(C_2) + P(A|C_3)P(C_3) + P(A|C_4)P(C_4) + P(A|C_5)P(C_5)$$
$$= \frac{P(\{1\})}{P(\{1,6\})} \bullet P(\{1,6\}) + \frac{P(\{2\})}{P(\{2,7\})} \bullet P(\{2,7\}) + \frac{P(\{3\})}{P(\{3,8\})} \bullet P(\{3,8\}) + \frac{P(\{4\})}{P(\{4,9\})} \bullet P(\{4,9\}) + \frac{P(\{5\})}{P(\{5,10\})} \bullet P(\{5,10\})$$
$$= P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) + P(\{5\})$$
$$= P(A)$$

Q6. (10 points)

(a) Let Ω = {a,b,c} be a sample space. Let P(a) = ½, P(b) = 1/3, and P(c) = 1/6. Find probabilities for all subsets of Ω.

Solution:

Possible subsets: {{a},{b},{c},{a,b},{a,c}, {b,c},{a,b,c},{NONE}}

Probability of each of them:

P(a) = ½

P(b) = 1/3

P(c) = 1/6

P(a,b) = ½ + 1/3 = 5/6

P(a,c) = ½ + 1/6 = 2/3

P(b,c) = 1/3 + 1/6 = ½

P(a,b,c) = ½ + 1/3 + 1/6 = 1

P(NONE) = 0

(b) The following table summarizes the results of breathing test given to drivers suspected of driving under influence. Suppose if two persons included in the following table are randomly selected without replacement, the find the probability that the first person has positive test result and the second person has negative test result.

	Persons actually driving after consuming alcohol	Persons without alcohol consumption
Positive test result	90	10
Negative test result	5	95

Solution:

P(first person = positive, second person = negative) =

$$\frac{90+10}{(90+10+5+95)} \times \frac{5+95}{99+5+95} = \frac{50}{199} \approx 0.2513$$