CSS 591 : Foundations of Data Science HW1

Name : Shashank Shekhar
Student Id : 2002623 27

Ans1- Various forms of data preprocessing are :

1) Data cleaning → This involves removing of outliers, clearing up noisy data and filling in missing values.
2) Data integration → In this step, all the data is combined.
3) Data Transformation → This involves the process to transform the data into reliable shape by using techniques such as normalization and aggregation.
4) Data Reduction → This step involves reducing the data volume but using this reduced data volume to generate similar analytical results.
5) Data Discretization → The step involves putting the data values (for each attribute) into buckets so as to limit the number of possible states and thus enable the algorithms to produce a mining model.

Ans2-

1) Table : Cust.Survey

CId	Gender	E_Grocery	E_Softdrink	Pref_Softd	Feedb_Softd	CAge	CRent
C1	M	42.36	6.42	Coke	VS	21	VL
C2	F	31.38	3.48	Pepsi	S	25	L
C3	F	18.09	10.00	Diet Coke	NS	33	NL
C4	M	16.00	2.38	Dr. Pepper	NS	42	NL
C5	M	23.50	11.45	Coke	S	18	L

2) ① CId (Customer Id) → Nominal Attribute

There is no specific meaning of the Id's except for the fact that they are used to differentiate between the customers. There is no specific order that makes one greater than the other.

② Grender → Nominal Attribute

The two categories, ie. male & female are used to differentiate between the customers and put them in 2 seperate groups with no comparison between the groups.

③ E_Grocery (Experse on Grocery) → Ratio Attribute

This is ratio because the expense values can be compared, have a fixed interval and have an absolute zero value ie. 0 which means no expense.
eg:- Expense of $2 is half of $4 expense.

④ E_softdrink (Expense on Softdrink) → Ratio Attribute

This is ratio because the expense values can be compared, have a fixed interval and have an absolute zero value.

⑤ Pref_Softdr (Preferred Softdrink) → Nominal Attribute

This attribute is nominal because preferrence of the customers cannot be compared -ie.

one is better than the other.

(6) Fcedba Softd (Feedback for softdrink) → Ordinal Attribute

This is because we can compare between different customer feedbacks. This is not interval because the difference between the different values for this attribute is not quantified and hence not fixed.

Not Satisfied < Satisfied < Very Satisfied.

(7) CAge (Customer Age) → Ratio Attribute

This is because the age is a number that can be used to differentiate customers plus follows fixed intervals meaning (22-21)=(21-20) and ratios have meanings. Eg:- C1 = 20 yrs and C2 = 40 yrs then we can say that C2 = 2 times of C1 and also absolute zero is defined meaning age of zero means person isn't born.

(8) CRet (Customer Retention) → Ordinal Attribute

This is because we can compare between different values as

Very Unlikely < Not Likely < Likely < Very Likely.

3) Statistics for the attributes:

(1) CId (Customer Id) → Count

Count is meaningful since it gives the total count of the customers.

(2) Gender → Mode

Mode is meaningful since it will tell which of male or female has higher occurance (frequency).

(3) E_Grocery (Expense on Grocery) → Since all mathematical operations are allowed, the geometric mean, harmonic mean including mean, median and mode can summarize this attribute.
Median can be used when there are outliers and mean can be used when there are no outliers.

(4) E_Softd (Expense on Softdrink) → Median, Mean, Mode, Geometric Mean and Harmonic Mean

All the above can be used to summarize this attribute. Mean and Median gives the measure of central tendency. Mode gives the frequency of the values. Geometric and harmonic mean make sense since taking ratio is a valid operation.

⑤ Pref-Softd (Preferred Softdrink) → Count, Mode

Mode → This will give the highest preferred soft drink among the people

Count → This will reflect the counts of various types of softdrinks that people have.

⑥ Feedb-Softd (Feedback for Softdrink) → Median, Mode

Since the values ~~may contain~~ cannot be quantified in terms of the interval spacing, mean is not allowed whereas median (middle ranked element) is allowed to measure central tendency.

Mde is allowed to understand the highest frequency occuring element.

⑦ CAge (Customer Age) → Since all mathematical operations are allowed, the geometric mean, harmonic mean including mean, median and mode are allowed as measure of central tendency.

⑧ CRent (Customer Retention) → Median, Mode

Median is used to measure the central tendency.
Mode is used to find the frequency of the highest occuring value.

4) Graphical representation for each use case:

(1) CId (Customer Id) → Bar Chart, Pie Chart

Because the values are discrete with no intervals, bar and pie chart are the options for graphical representation.

(2) Gender → Bar Chart, Pie Chart

Since the values are discrete with no proper interval defined, bar and pie charts are the only two options available.

(3) E_Grocery (Experse on Grocery) → Since this a continuous attribute with a range defined and intervals, histogram and box plots are used.

(4) E_Softdrink (Experse on Softdrink) → Histogram and Boxplot

Since this is also a continous attribute with defined intervals, histogram and box plots are used.

(5) Pref_Softd (Preferred Softdrink) → Bar Chart, Pie Chart

These are discrete values with no correlation between them. Hence, bar chart and pie chart are preferred.

⑥ Feedb_Softd (Feedback for softdrinks) → Bar Chart and Pie Chart

Since the attribute holds discrete values, bar chart and Pie Chart are used.

⑦ CAge (Customer Age) → Histogram and Boxplot.

Since age is defined as a ratio attribute with a range (divided into interval), histogram and boxplot are suitable.

⑧ CRent (Customer Retention) → Bar Chart and Pie Chart

Since the attribute holds discrete values, bar chart and Pie chart are used.

Ans 3-

a) Statistics → This is the branch of mathematics that deals with the collection, analysis, interpretation and presentation of numerical data.

b) Population → A set of events or similar items that are of interest for some experiment.

c) Sample → These are the set of observations drawn from the population.

d) Event Space → This contains all possible events for a given experiment.

e) Event → Some subset of outcomes from the sample space is called an event.

f) Random Variable → This is a variable that takes a value as the outcome of a statistical experiment

g) Experiment → Any procedure that can be infinitely repeated and has a well-defined set of possible outcomes

h) Discrete Data → A data set that is countable and takes only a finite number of values. eg:- number of apples in basket.

i) Continuous Data → A data set that can take any value within a range, eg:- Temperature readings in a range.

j) Mean → It is a measure of central tendency obtained by adding all the values in a data set and then dividing by the total number of values that are added.

k) Median → It is a measure of central tendency. For an odd number of observations, it is obtained by arranging the elements in ascending order and then the middle value is the median. For an even number of observations, it is the mean of the two middle numbers when the numbers are arranged in increasing order

l) Variance → It is the average of the squared differences from the mean.

Ans 4- Data Points:

98.8, 98.4, 98.2, 98.1, 99.0, 98.9, 99.2, 98.3, 98.2, 98.8

(μ) Mean $= \frac{\sum_{i=1}^{n} d_i}{n}$

$$\mu = \frac{985.9}{10} = 98.59$$

$$\text{Variance} = \frac{(98.8-98.59)^2 + (98.4-98.59)^2 + (98.2-98.59)^2 + (98.1-98.59)^2 + (99-98.59)^2 + (98.9-98.59)^2 + (99.2-98.59)^2 + (98.3-98.59)^2 + (98.2-98.59)^2 + (98.8-98.59)^2}{10}$$

$$\text{Variance} = \frac{0.0441 + 0.0361 + 0.1521 + 0.2401 + 0.1681 + 0.0961 + 0.3721 + 0.0841 + 0.1521 + 0.0441}{10}$$

$$\text{Variance} = 0.1389$$

$$S.D(\sigma) = \sqrt{\text{Var}} = \sqrt{0.1389} = 0.37269$$

Outliers $= \mu \pm 2\sigma$

$= 98.59 + (2 \times 0.37269)$; $98.59 - (2 \times 0.37269)$

$=> 99.3354$; < 97.8446

i) Values greater than 99.3354 and less than 97.8446 are Anomalies. 104.2°F is beyond this and hence it is an anomaly.

ii) Yes, in general medical terms, 104.2° F is the fever (illness) temperature. Yes, the definition of anomaly defined is usual in this because temperatures beyond the normal region, (97.8446 < temp < 99.3354), are the temperatures where the people fall ill.

iii) To find out the z-score :

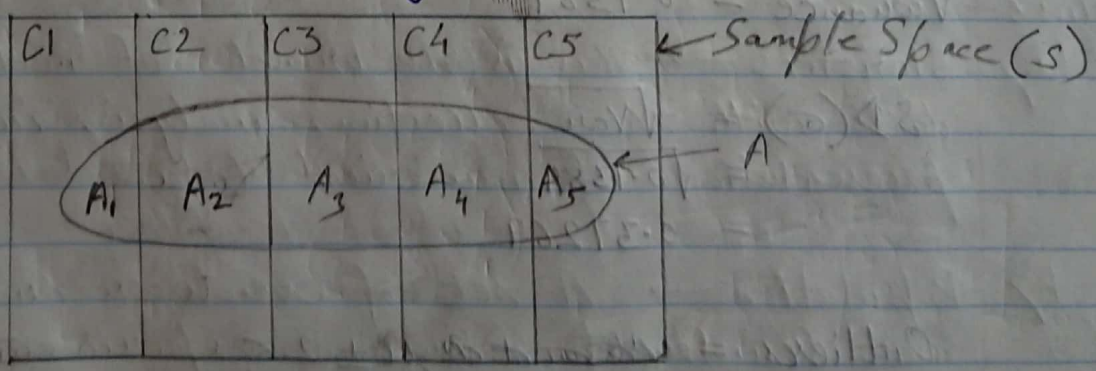
$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{104.2 - 98.59}{0.37269}$$

$$= 15.05$$

The reading of 104.2°F is 15.05 standard deviations above the mean.

Ans5- Law of total probability using Venn Diagram.



Probability of A, $P(A) = P(A_1) + P(A_2) + P(A_3) + P(A_4) + P(A_5)$

Since $A_1 \subseteq B$, hence $P(A_1) = P(A_1 \cap C_1) = P(A_1 | C_1) P(C_1)$

$P(A) = P(A \cap S)$, where S represents Sample Space

Also, given $C_1 \cup C_2 \cup C_3 \cup C_4 \cup C_5 = S$

Hence, $P(A) = P(A \cap (C_1 \cup C_2 \cup C_3 \cup C_4 \cup C_5))$

$= P(A \cap C_1) \cup P(A \cap C_2) \cup P(A \cap C_3) \cup P(A \cap C_4) \cup P(A \cap C_5)$

From Probability Axiom, $P(A \cap B) = P(A|B) \cdot P(B)$ and hence

$$P(A) = P(A|C_1) \cdot P(C_1) + P(A|C_2) \cdot P(C_2) + P(A|C_3) \cdot P(C_3) + P(A|C_4) \cdot P(C_4) + P(A|C_5) \cdot P(C_5)$$

Hence, proved.

Ans 6 - a) $\Omega = \{a, b, c\}$

Event Space,

$E_1 = \{\phi\} \Rightarrow P(E_1) = 0$

$E_2 = \{a\} \Rightarrow P(E_2) = 1/2$

$E_3 = \{b\} \Rightarrow P(E_3) = 1/3$

$E_4 = \{c\} \Rightarrow P(E_4) = 1/6$

$E_5 = \{a, b\} \Rightarrow P(E_5) = \frac{1}{2} \times \frac{1}{3} = 1/6$

$E_6 = \{b, c\} \Rightarrow P(E_6) = \frac{1}{3} \times \frac{1}{6} = 1/18$

$E_7 = \{a, c\} \Rightarrow P(E_7) = \frac{1}{2} \times 1/6 = 1/12$

$E_8 = \{a, b, c\} \Rightarrow P(E_8) = \frac{1}{2} \times \frac{1}{3} \times \frac{1}{16} = \frac{1}{96}$

b) Let A = event that 1st person has +ive test result and B = event that 2nd person has -ive test result.

$$P(A \cap B) = \frac{100}{200} \times \frac{100}{199}$$

$$P(A \cap B) = \frac{50}{199}$$