Answer 1)

Data preprocessing is defined as the process where the collected data is transformed into understandable format and all the discrepancies are removed. In practical data analysis, the volume of data is large and the data collected may have errors. If this data is directly used for representation, it will lead us to a false inference and a wrong choice of action. This can be fatal in critical situations such as medical diagnosis and emergency response systems. Hence data preprocessing is an important step in data analysis.

The data which is collected and not preprocessed is called as raw data. There is a high chance for the raw data to be incomplete, noisy and inconsistent. Incomplete data can be a result of absence of the required attributes while collecting data. Noisy data refers to correct attribute values which maybe a result of malfunctioning instruments or fault in transferring the recorded values. Inconsistent data may occur if a certain value is deleted from one place and not updated in other places of occurence. There are various steps involved in data preprocessing as follows:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

DATA CLEANING:

The three characteristics of the raw data that were discussed above are removed in this process.

- Incomplete Data: The missing values for each attribute can be filled in using different ways.
    1. The values could be entered directly manually. This is a simple approach is not feasible in large databases and is not used in practice.
    2. Ignoring the missing values of attributes for that particular tuple. This is the simplest method to handle missing data but is not effective when there is a large volume of missing data or if the missing data directly affects the result.
    3. Using a common value or global constant in place of the missing values. Negative infinity($-\infty$) is commonly used in this case.
    4. Using the average of the attributes
    5. From the information available, the most probable value is calculated using the Bayesian formula
- Noisy Data: It consists of random errors and incorrect values. The data is smoothed and outliers are removed using the following techniques:
    1. Binning: The values of each attribute are sorted and distributed into bins. The distribution divides the attribute values into intervals with equal size or intervals with equal number of values.

2. Clustering: This techniques divides the values into groups or clusters. All the values belonging to one cluster have similar properties and characteristics. As a result of which outliers are detected and removed
3. Regression: The data is smoothed by applying linear regression or multiple regression functions. In linear regression a straight line is used to fit two variables. Once this is done, one variable can be used to predict another variable. Multiple regression is similar to linear regression, but the number of variables are more than two.

- Inconsistent data: The inconsistencies are removed manually by referring to the external or dependent data. In accordance with that, knowledge engineering tools are used to enforce data constraints.

DATA INTEGRATION

While collecting data, various sources are used which collects data and stores them into different files and formats such as flat files, relational databases etc. The sources of data may include different database management systems. Data integration is the process of combining the data from various sources into a single data store. While integrating the data, two values could be duplicated or repeated, in such a case redundancy occurs. This process of data integration removes all the redundant values. The naming of the attributes is generalized to remove any duplications

DATA TRANSFORMATION

The data is transformed into appropriate type using various techniques:

- Normalization: The data values are expanded or reduced so as to fall within a certain interval/range. The scaling is done on the entire dataset so as to avoid anomalies
- Aggregation: This is the process of summarizing the data values into an attribute value f a higher level in the hierarchy. For example: the values of days can be aggregated into weeks/ months/ years
- Generalization: It is opposite to aggregation, where the values are further categorized to a lower level of hierarchy. For example, The numerical attributes of employees at a college can be further divided into teaching faculty, administration, finance.

DATA REDUCTION

The entire dataset consists of many values, all of which are not required while performing a query or analysis. The values which are required are only used for computation which improves the complexity of the process. Hence, for implementing this approach, particular data values are selected, thereby reducing the dataset to work on. This process is called as data reduction. The different strategies for data reduction are:

- Reducing the number of attributes:
    - Operations such as slice, dice and roll-up is used to select the appropriate tuples from the dataset to work upon.

    - The irrelevant attributes are removed from the dataset
    - Algorithm such as Principal Component analysis is performed on the dataset to reduce the dataset from say three attributes to two attributes.
- Reducing the number of attribute values:
    - Binning: As discussed above, in this process, the values of a single vector are used to represent a certain axis is divided into intervals, thereby reducing the number of values on that respective axis.
    - Clustering: the data items are grouped together into clusters who have the same properties
    - Aggregation and Generalization: This reduces the dataset by combining or collapsing the data items as necessary.

To summarize, data processing is an important step before making any conclusions from the dataset as it can lead to false results. Data processing helps us to save computing time and increases the efficiency. It removes all the abnormalities in data and prepares the data to be worked on and infer the results accordingly.

References: (https://www.xenonstack.com/blog/data-science/preparation-wrangling-machine-learning-deep/
http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-3.html
https://www.techopedia.com/definition/14650/data-preprocessing)

Answer 2)

1)

Customer_id	**Gender**	**Age**	**Amount_spent_groceries**	**Amount_spent_beverages**	**Soft_beverage**	**Satisfaction**	**Will_buy**
sb1	M	22	160	25	Pepsi	Satisfied	Likely
sb2	F	36	175	45	Coke	Not Satisfied	Very unlikely
sb3	F	19	120	38	Dr. Pepper	Satisfied	Likely
sb4	M	40	135	26	Coke	Very Satisfied	Very likely

2)

Sr.No	Attribute	Data Type	Reason
1	Customer_id	Nominal	Customer_id can only map a person to the respective id. We are not able to perform any other mathematical operations on this data
2	Gender	Nominal	This attribute is a qualitative measure and only provides two choices. It can only be used to distinguish whether a particular sample is male or female. None of the mathematical operations can be performed on it
3	Age	Ratio	It is a quantitative measure of data and all types of geometric operations can be performed on it. The scale for age has a fixed zero point.
4	Amount_spent_groceries	Ratio	This attribute can never have a negative value as the amount spent can never be less than zero. Since the position of zero is defined, it is a ratio
5	Amount_spent_beverages	Ratio	Similar to the above reason, the amount spent on beverages can only be greater than or equal to zero. Here, the location of zero is fixed too

6	Soft_beverage	Nominal	The type of soft beverage can be anything which belongs to the set of soft beverages. The values cannot be mathematically compared to each other.
7	Satisfaction	Ordinal	This is a qualitative measure and there can be a comparison between the different values of this attribute which belong to a finite set (Very satisfied, Satisfied, Not Satisfied), hence this qualifies as an ordinal data type
8	Will_buy	Ordinal	It is a qualitative measure where the values belong to an ordered set (Very likely, Likely, Not Likely, Very unlikely). These categories follow a distinctive order of decreasing likeness hence, this can be considered as an ordinal data type

3)

Sr.No	**Attribute**	**Summary statistics**	**Reason**

1	Gender	Mode	Since gender is of a ordinal data type, the only statistic we can use is mode where we are able to show the the most frequent value which occurs. We can find which gender has given the maximum number of responses
2	Age	Median, inter-quartile range(IQR)	The mean helps us to find the rough age of all the participants in the survey. There arises a problem when a person belonging to an extreme age (very old/ very young) brings up or brings down the average of the entire set. The median and IQR helps us to resolve this issue by giving the age of the sample at a quarterly interval. It helps us to understand the measure of spread
3	Amount_spent_groceries	Mean, standard deviation	The mean helps to find out the rough estimate about the amount spent by each person in the entire sample, while the standard deviation helps to find how the amount spent on groceries is spread around the mean.This will help us to understand the expenditure of an average user on food items which can be further used to

			compare it with the amount spent on beverages.
4	Amount_spent_beverages	Mean, standard deviation	By calculating the mean and standard deviation, we can estimate the amount spent by a single person on soft drinks and by this information conclude the amount which is to be labelled for the new product (diet coke) or derive other possible behaviours of the customers
5	Soft_beverage	Mode	The mode will help us to understand the most preferred soft drink by the entire population which helps us to understand the current trend
6	Satisfaction	Mode	Calculating the mode helps us to understand the if people are satisfied or not. If the mode gives us a positive trend, then the diet coke is more likely to sell faster, otherwise additional steps would be required to sell the product
7	Will_buy	Mode	This is the most important attribute which will help us to predict whether people are willing to buy the beverage or not. The mode will give the most opted answer by the people

			and helps us to conclude whether a majority of people are willing to buy diet coke or not

4)

Sr No	Attribute	Graphical Representation	Reason
1	Gender	Pie chart	It will help us to understand the percentage of people from the sample that are either male or female
2	Age	Histogram	It will help us divide the ages into intervals and directly compare the number of people in each interval
3	Amount_spent_groceries	Line Chart	It will help us to describe the trend of the expenditure on groceries and easily figure out the amount that is spent the most by the people
4	Amount_spent_beverages	Line Chart	It also helps to understand the tendency of the people to spend money on soft beverages. By directly looking at the line graph, we can conclude what is the amount that people would be willing to spend on soft beverages
5	Soft_beverage	Pie Chart	This can be best described by pie

			chart which will be able to show the percentage of each soft drink preferred by the entire sample population
6	Satisfaction	Histogram	Histogram will show us the frequency of people for each category, thereby graphically representing the most favourable and the least favourable result
7	Will_buy	Histogram	The will_buy histogram will show us the preference of the sample set for each category, thereby helping us to easily conclude whether many people will buy the 6 pack diet coke or not

References: (Birger Stjernholm Madsen: Statistics for Non-Statisticians, Springer)

Answer 3)

a) **Statistics**: It is defined as the techniques used for collecting data, analyzing the collected data and finally representing data in a concise form.
b) **Population**: The entire set of individuals on which the survey is performed to collect information is called as a population.
c) **Sample**: It is a subset of population of a survey which is randomly selected to analysis and graphical representation.
d) **Event Space**: It is defined as a set that contains all the possible outcomes of an experiment. Example: the event space for tossing a coin is {Heads, Tails}
e) **Event**: It is a subset of event space. The probability is calculated for all such events. Example:The outcome being heads after tossing a coin
f) **Random Variable**: It is a numerical value which is assigned to each outcome of an experiment. Example: We can consider the random variable X=0 (If the outcome of flipping a coin is Heads or X=1 if the outcome is tails)
g) **Experiment**: The activity that is performed to collect information about the data which is not readily available or that cannot be made available. The probability calculated for the

experiments is not the same as theoretical probability. It is only true if the number of samples is large

h) **Discrete Data**: It consists of whole quantity that can be counted and can be an infinite in number. Example: The set of natural numbers
i) **Continuous Data**: It can be defined as the data where there is no interval/gap between two values. The continuous data values are defined in a range. Example: Floating point data type
j) **Mean**: It is the measure of location of any given attribute of the data set. It is the average of the entire attribute and is calculated as (Total Value of the given attribute/ No of values for that attribute)
k) **Median**: It is defined as the center of the sample space. Once the data is sorted, the value at the center of the sample space is the median. It is calculated as (Number of values/2 ; if the number of values is even and average of the two middle values ; if the number of values is odd)
l) **Variance**: It is defined as the measurement of spread of data from the mean value. In other words, it tells us how far a particular value of the attribute is from its mean. It is calculated by squaring the difference of the individual values from the mean.

References: (Birger Stjernholm Madsen. Statistics for Non-Statisticians, Springer/ Géza Schay. Introduction to Probability with Statistical Applications, Springer)

Answer 4)
i)
Mean = 98.8 +98.4 +98.2 +98.1 +99.0 +98.9 +99.2 +98.3 +98.2 +98.8/ 10

Mean= 98.59

Value	X= Value- Mean	Y= (Value-Mean)^2
98.8	0.21	0.0441
98.4	-0.19	0.0361
98.2	-0.39	0.1521
98.1	-0.49	0.2401
99.0	0.41	0.1681

98.9	0.31	0.0961
99.2	0.61	0.3721
98.3	-0.29	0.0841
98.2	-0.39	0.1521
98.8	0.21	0.0441

Now we calculate the mean of Y:

Mean= 0.0441+ 0.0361+ 0.1521+ 0.2401+ 0.2401+ 0.0961+ 0.3721+ 0.0841+ 0.1521+ 0.0441/10

Mean= 0.1389

Now, we find the square root of this calculated mean to get the standard deviation:

Standard Deviation= 0.3727 ~ 0.37

To find whether temperature of 104.2 deg F is an anomaly we check if it is greater than 2 standard deviations:

Therefore,

Original Mean + 2 standard deviation= 98.59+ 2* 0.37
= 99.33 deg F

Hence, we conclude that the body temperature of 104.2 deg F is unusual as it lies more than 2 standard deviations away from the mean.

ii)

The normal body temperature of a person lies between 97-98 deg F which may reach to 100-100.5 deg F in cases of normal fever. Hence, by general medical knowledge, the body temperature of 104.2 deg F is far away from normal.

Here, the mean body temperature of patients was 98.59 deg F which states that most of the data is surrounding the 98.59 degF. In such a case, a value of 104.2 deg F is at a larger distance as compared to the other values in the data set which the maximum being 99.2. As 104.2 is at a greater distance from 99.2 than 99.2 deg F is from 98.59 deg F, we can easily conclude that 104.2 deg F is unusual without the need of the definition of anomaly.

iii)

z score/ standardized score of a value is defined as the number of standard deviations away from the mean.

Z score for 104.2 deg F is calculated as follows:

z-score= value- mean/ standard deviation

z-score= 104.2- 98.59/ 0.3727

z-score= 15.05

Hence, the z-score for 104.2 deg F is 15.05

Answer 5)

The law of conditional probability gives us the probability of an event say A such that event B has already occured before. The formula is given as follows:

P(A)= P(A|B)P(B)

Consider a question where we roll a fair die and we need to find the probability of an even number, given that the result is less than 5.

We solve it by first finding the probability of an even number. The sample set contains {2, 4, 6}

Hence, P(A)= |{2,4,6}| / 6 = ½

Since we know that the outcome is a number less than equal to 5 (Event B), we can reduce our sample space to {2, 5}. Here we are considering both the events A and B, therefore we are finding the probability of A given B (A | B)

P(A | B)= | A ∩ B | / | B | = |{2,4}| / |{1,2,3,4,5}| = ⅖

So we can say: P(A | B)= | A ∩ B | / | B |

To generalize, we can divide the RHS by sample space |S| ,hence we get:
P(A | B)= (| A ∩ B |)/ (| B | / |S|)

P(A | B)= P(A ∩ B) / P(B)

We can derive a formula that **P(A ∩ B) = P(A | B) * P(B)**

Consider the given question to prove the law of total probability:

To prove:
P(A) = P(A | C1)P(C1) + P(A | C2)P(C2) + …+P(A| Cm)P(Cm)

If we consider m to be 5 then we have to prove the following:

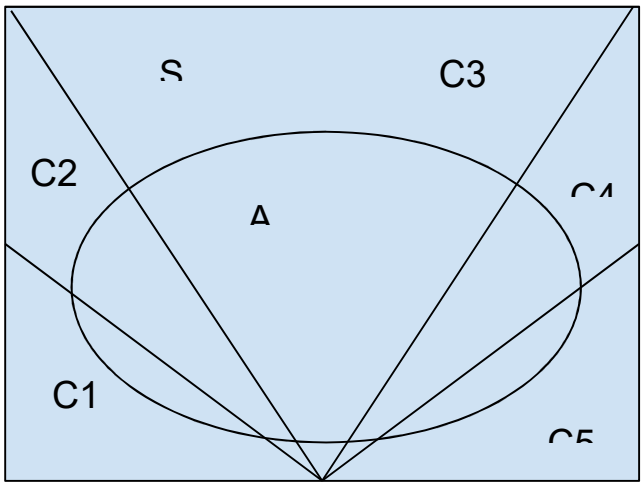
P(A)= P(A | C1)P(C1) +P(A | C2)P(C2) +P(A | C3)P(C3) +P(A | C4)P(C4) +P(A | C5)P(C5)

Using the law of conditional probability we can simplify the equation as:

P(A)= P(A ∩ C1)+ P(A ∩ C2)+ P(A ∩ C3)+ P(A ∩ C4)+ P(A ∩ C5)

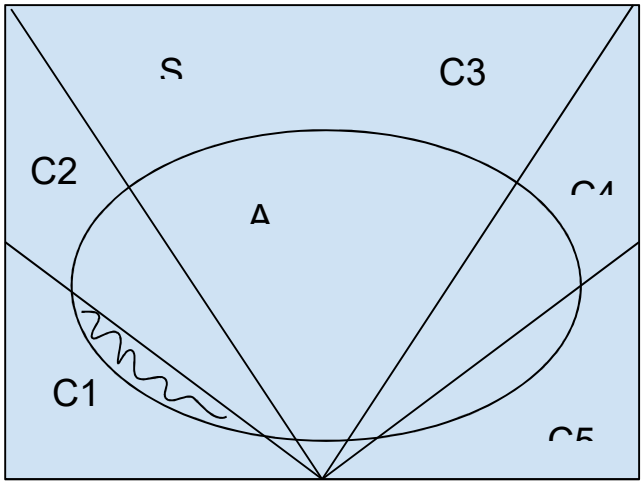
Let us prove this using a venn diagram:

Consider an sample S. As given, there are five disjoint events C1, C2, C3, C4, C5 that make up the sample space. Another arbitrary event A occurs as well. It can be shown as follows

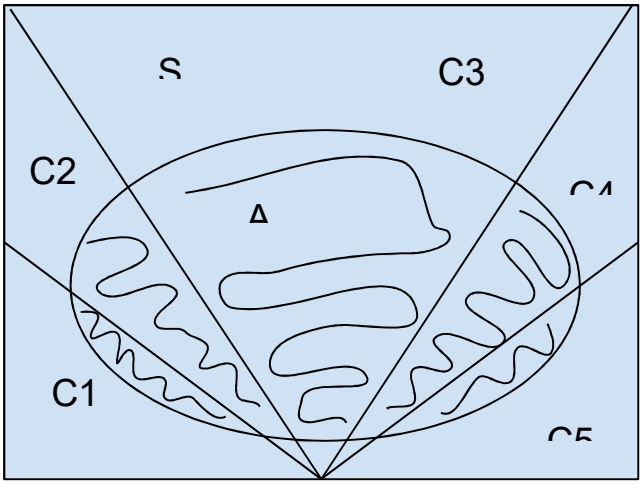


Since events C1...C5 are disjoint they do not overlap each other and C1+C2+C3+C4+C5 = S
It shows that A is made up of different sections which coincide with the sections C1...C5

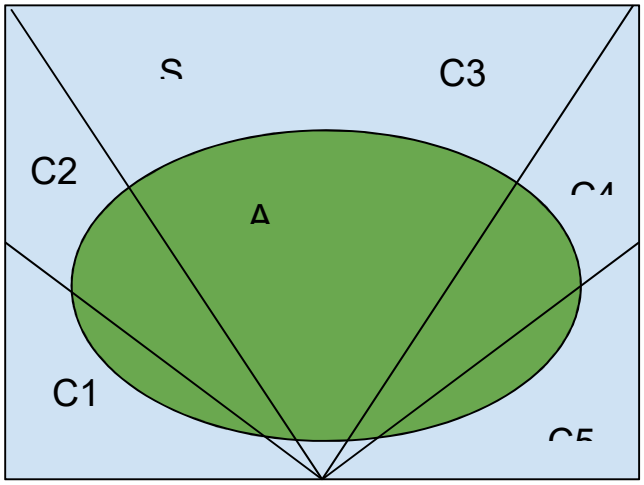
Now, A ∩ C1 can be shown as :



Similarly we can show A ∩ C2, A ∩ C3, A ∩ C4, A ∩ C5



Now if we add up all the regions we get the following shaded region which is equal to A



This proves that :

$|A \cap C1| + |A \cap C2| + |A \cap C3| + |A \cap C4| + |A \cap C5| = A$

Now if we divide both the sides by sample space S, we get:

$P(A \cap C1) + P(A \cap C2) + P(A \cap C3) + P(A \cap C4) + P(A \cap C5) = P(A)$

This can be also represented by using the conditional probability as:
$P(A) = P(A \mid C1)P(C1) + P(A \mid C2)P(C2) + P(A \mid C3)P(C3) + P(A \mid C4)P(C4) + P(A \mid C5)P(C5)$

In this way, we have proven the law of total probability using the venn diagram.

References: (Birger Stjernholm Madsen. Statistics for Non-Statisticians, Springer/ Géza Schay. Introduction to Probability with Statistical Applications, Springer)

Answer 6)
a)

Ω= {a,b,c} ; All the possible subsets are:

Subset	Probability	Reason
ϕ	0	After every event, there is going to be an outcome. Hence the probability of the outcome being null is always zero
{a}	½	Given
{b}	⅓	Given
{c}	⅙	Given
{a,b}	⅚	P(a∪b)= P(a)+P(b)= ½+⅓= ⅚ (Axiom 3)
{a,c}	⅔	P(a∪c)= P(a)+P(c)= ½+⅙ = ⅔ (Axiom 3)
{b,c}	½	P(b∪c)= P(b)+P(c)=⅓+ ⅙ (Axiom 3)
{a,b,c}	1	When any event takes place the probability of the outcome belonging to the set of samples is always going to be 1 as the sample space always contains all the outcomes. (Axiom 2)

b)

Total number of people driving the car= 90+10+5+95= 200

Number of positive test results= 90+10= 100

Therefore, the probability that the first person has positive test result is:

P1= 100/200= 0.5
Hence, probability that the first person has positive test result is 0.5

Now, from the remaining 199 people (since there is no replacement), the probability that the second person has negative test result is:

P2= Number of negative test results/ Total number of people driving the car= 95+5/199=

100/199= 0.503

Hence, probability that the second person has negative test result is 0.503

Therefore the final probability is:
P= P1*P2= 0.2513

The probability that the first person has positive test result and the second person has a negative test result is 0.2513

References: (Géza Schay. Introduction to Probability with Statistical Applications, Springer)