Class X

ANSWER KEY

Data Science

1. Use of Statistics in Data Science

Unsolved Exercise

Objective Type Questions

A. 1. d 2. c 3. c 4. a 5. c

6. d 7. b

8. This question is misprint in the book. Please replace it with the following question:

Find the population variance of the given dataset: 3, 9, 5, 6, 7.

a. 1 b. 2

c. 3 d. 4

Ans. d

9. a 10. c

B. 1. Subset 2. Frequency 3. mode 4. low 5. mean

6. empirical relationship 7. Mean Absolute Deviation 8. median

9. central tendency 10. Tally marks

Standard Questions

A. 1. Central tendency is also known as average.

2. mean, mode and median

3. Median should be used over mean when there are some irregular values also known as outliers in the dataset.

4. In the case of a moderately skewed distribution, i.e., in general, the difference between mean and mode is equal to three times the difference between the mean and median.

Thus, the empirical relationship as Mean – Mode = 3 (Mean – Median)

5. In the case of a frequency distribution, all the three central tendency values, i.e., mean, mode and median are equal:

   mean = median = mode

6. Dispersion refers to the extent of variation or spread between the various values of a dataset.

7. The term "standard deviation" refers to a measurement of the data's dispersion from the mean. A low standard deviation means that the data are grouped around the mean, whereas a high standard deviation means that the data are more dispersed.

8. The first 5 prime numbers are: 2, 3, 5, 7, and 11.

   Mean = $(2 + 3 + 5 + 7 + 11) / 5$

   $= 28 / 5$

   $= 5.6$

9. $8+11+6+14+x+13=11\times6$

   $8+11+6+14+x+13=66$

   $52+x=66$

   $x=66-52$

   $x=14$

   So, the value of the observation x is 14.

10. Mean = $(6+8+(x+2)+10+(2x-1)+2)/6$

    $9 = (6+8+(x+2)+10+(2x-1)+2)/6$

    $54 = 6+8+x+2+10+2x-1+2$

    $54 = 18+3x$

    $3x = 54-18$

    $x = 36/3$

    $x = 12$

    Now, substituting $x = 12$ into one of the expressions, say $2x-1$, we find:

    $= 2(12) - 1$

    $= 24 - 1$

    $= 23$

    So, the value of x is 12, and the value of the observation in the data is 23.

11. Given the weights: 39, 43, 36, 38, 46, 51, 33, 44, 44, 43.

    In this case, both 43 and 44 occur twice, making them the modes.

    So, the mode(s) of the data set are 43 and 44.

12. First ten whole numbers are 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10.

    Sum of the first ten whole numbers: $1+2+3+4+5+6+7+8+9+10=55$

Total count of numbers: 10

Mean = Sum of numbers / Total count of numbers

Mean = 55/10

Mean = 5.5

B. 1. There are many methods of subsetting the data, three of which are discussed below:

- Row-based subsetting
- Column-based subsetting
- Data-based subsetting

Row-based Subsetting

In this method of subsetting, we take some rows from the top or bottom of the table For example, you may need to subset the rows of a data frame because you may be interested in understanding a subpopulation in given sample.

Column-based Subsetting

Occasionally the original dataset may contain a large number of columns and all of them may not be essential to perform the analysis. You may then choose specific columns from the dataset. This process of subsetting is known as column-based subsetting.

Data-based Subsetting

To subset the data based on specific data we use data-based subsetting. Data-based subsetting is creating a copy of a database that contains only a portion of the data, based on certain criteria while still being referentially intact.

2. The two-way frequency table and the two-way relative frequency table are extremely similar. The main difference is that, in case of two-way relative frequency table, we take percentages rather than numbers. Tables of two-way relative frequency show the percentage of data points that fall into each group. Depending on the context of the issue, you can use either row relative frequencies or column relative frequencies.

3. Generally, mean and median both represent the central tendency of a dataset. So when should we use median over mean? Median is a more accurate form of central tendency especially when there are some irregular values also known as outliers.

For example, consider the given situation. Your uncle gets his blood pressure checked every week. But due to some error in the device, the recording for one week was too high.

140, 142, 145, 220, 147

Here, 220 is the error recorded by the instrument and considered to be an outlier. In this case, the mean is 158.8 and the median is 145. However, the median value may also give the wrong signal, as in the worst scenario, the median value is itself an outlier, as in the given scene.

Due to this simple error of the device, the mean value deviates greatly from the regular blood pressure values due.

Whereas the median value still correctly represents the central point of the dataset. Thus, under conditions where there are outliers in the dataset, the median is a more effective measure of central tendency.

But, here, in the above given data, the mean is same and median may be very high. In all such cases the outlier must be discarded.

Also, the median is preferred, especially when the data distribution contains some extremely low and high values. In these circumstances, the median is a more accurate measure of central tendency than the mean. The median is typically preferred over the mean when determining compensation for the simple reason that the median is far less impacted by outliers (abnormally low or high numbers) than the mean.

4. This question is misprinted in the book. Please replace it with the following question:

   What is the purpose of subsetting?

   Ans. Subsetting the data is a useful indexing feature for accessing object elements. It can be used for selecting and filtering variables and observations. We subset the data from a data frame to retrieve a part of the data that we need for a specific purpose. This helps us observe just the required set of data by filtering out unnecessary content. For example, Subsetting allows you to work with data that contains all the necessary links between tables for your programs to function, but for a fraction of the cost.

   A subset allows you to reduce the size of your dataset, and to split the examples into disjoint sets for the purpose of training, validation, and testing. In research communities (for example, earth sciences, astronomy, business, and government), subsetting is the process of retrieving just the parts of large files which are of interest for a specific purpose. This occurs usually in a client-server setting, where the extraction of the parts of interest occurs on the server before the data is sent to the client over a network. The main purpose of subsetting is to save bandwidth on the network and storage space on the client computer.

5. Standard deviation is calculated as follows:
   - Find the mean by adding up all the data parts and dividing it by the number of parts of the data.
   - Subtract mean from each value.
   - Calculate the square of each of the differences.
   - Find the average of squared numbers calculated in previous point to find the variance.
   - Finally, find the square root of the variance. That is the standard deviation.

6. To find the variance:
   - Find the mean by adding up all the data parts and dividing it by the number of parts of the data.
   - Subtract mean from each value.
   - Calculate the square of each of the differences.
   - Find the average of squared numbers calculated in previous point to find the variance.

7. Mean=(117+16+121+51+101+81+1+16+9+11+16)/11

   =430/11

   Mean≈39.09

   To find the mode, we count the frequency of each run:

   1 occurs once

   7 occurs once

9 occurs once

11 occurs once

16 occurs three times

51 occurs once

81 occurs once

101 occurs once

121 occurs once

The mode is 16 because it appears the most times (three times).

First, we need to sort the data set:

1,7,9,11,16,16,16,51,81,101,121

Since the number of observations is odd (11), the median is the middle value, which is the sixth value in the sorted list: Median=16

8. a. For the data set: 16, 18, 19, 21, 23, 23, 27, 29, 29, 35

   Sum of the data = 16+18+19+21+23+23+27+29+29+35=240

   Total number of values = 10

   Mean = 240/10=24

   So, the mean of the first data set is 24.

   b. For the data set: 2.2, 10.2, 14.7, 5.9, 4.9, 11.1, 10.5

   Sum of the data = 2.2+10.2+14.7+5.9+4.9+11.1+10.5=59.5

   Total number of values = 7

   Mean = 59.5/7≈8.50

   So, the mean of the second data set is approximately 8.50.

   c. Given data: 11/4, 21/2, 51/2, 31/4, 21/2

   To add fractions, we need a common denominator, which in this case is 4:

   = 11/4+42/4+102/4+31/4+42/4

   = (11+42+102+31+42)/4

   = 228/4

   = 57

   Now, since we have 5 values in the data set, we divide the sum by 5 to find the mean:

   Mean = 57/5

   =11.4

   So, the mean of the given data is 11.4.

9. a. Data set: 27, 39, 49, 20, 21, 28, 38

   Arranging the data in ascending order: 20, 21, 27, 28, 38, 39, 49

   Since the number of observations is odd (7), the median is the middle value, which is the fourth value in the sorted list: 28

   So, the median of data set (a) is 28.

   b. Data set: 10, 19, 54, 80, 15, 16

   Arranging the data in ascending order: 10, 15, 16, 19, 54, 80

   Since the number of observations is even (6), the median is the average of the two middle values, which are the third and fourth values in the sorted list: (16 + 19) / 2 = 35 / 2 = 17.5

So, the median of data set (b) is 17.5.

c. Data set: 47, 41, 52, 43, 56, 35, 49, 55, 42

Arranging the data in ascending order: 35, 41, 42, 43, 47, 49, 52, 55, 56

Since the number of observations is odd (9), the median is the middle value, which is the fifth value in the sorted list: 47

So, the median of data set (c) is 47.

d. Data set: 12, 17, 3, 14, 5, 8, 7, 15

Arranging the data in ascending order: 3, 5, 7, 8, 12, 14, 15, 17

Since the number of observations is even (8), the median is the average of the two middle values, which are the fourth and fifth values in the sorted list: (8 + 12) / 2 = 20 / 2 = 10

So, the median of data set (d) is 10.

10. a. Data set: 12, 8, 4, 8, 1, 8, 9, 11, 9, 10, 12, 8

The values and their frequencies are:

- 1 occurs once
- 4 occurs once
- 8 occurs four times
- 9 occurs twice
- 10 occurs once
- 11 occurs once
- 12 occurs twice

The mode is 8 because it appears the most times (four times).

So, the mode of data set (a) is 8.

b. Data set: 15, 22, 17, 19, 22, 17, 29, 24, 17, 15

The values and their frequencies are:

- 15 occurs twice
- 17 occurs three times
- 19 occurs once
- 22 occurs twice
- 24 occurs once
- 29 occurs once

The mode is 17 because it appears the most times (three times).

So, the mode of data set (b) is 17.

c. Data set: 0, 3, 2, 1, 3, 5, 4, 3, 42, 1, 2, 0

The values and their frequencies are:

- 0 occurs twice
- 1 occurs twice
- 2 occurs twice
- 3 occurs three times
- 4 occurs once
- 5 occurs once
- 42 occurs once

The mode is 3 because it appears the most times (three times).

So, the mode of data set (c) is 3.

d. Data set: 1, 7, 2, 4, 5, 9, 8, 3

The values and their frequencies are:

- 1 occurs once
- 2 occurs once
- 3 occurs once
- 4 occurs once
- 5 occurs once
- 7 occurs once
- 8 occurs once
- 9 occurs once

Since each value occurs only once, there is no mode.

So, the data set (d) is said to have no mode.

Higher Order Thinking Skills

1. Given dataset: [45, 78, 86, 48, 48, 75]

   Sum of the dataset = 45 + 78 + 86 + 48 + 48 + 75 = 380

   Total number of values = 6

   Mean = Sum of values / Total number of values Mean = 380 / 6 Mean = 63.33 (rounded to two decimal places)

   So, the mean of the given dataset is approximately 63.33.
2. Do it yourself.

2. Distributions in Data Science

Unsolved Exercise

Objective Type Questions

A.	1. a	2. b	3. a	4. c	5. d
	6. c	7. a	8. d		
B.	1. True	2. True	3. True	4. False	5. False
	6. True				

Standard Questions

A. 1. The Probability Density Function (PDF) P(x) of a continuous random variable X is defined as the derivative of the CDF P(x): P(x)=ddxFP(x).

2. A Binomial distribution is a common probability distribution that models the probability of obtaining one of two outcomes under a given number of parameters.

3. Collecting data helps answer the questions.

4. The formula for a discrete uniform distribution is:

$$Px = \frac{1}{n}$$

Where:

Px = Probability of a discrete value

n = Number of value in the range

5. PMF stands for Probability Mass Functions.

B. 1. A normal distribution is most common distribution function for independent, randomly generated variables. It is sometimes called the bell curve or Gaussian distribution, and is a distribution that occurs naturally in many real-life situations like IQ scores and represents natural phenomena such as errors, heights, weights, blood pressure, etc.

2. The term "variability" describes how dispersed a set of data is. In statistics, "variability" refers to the variation that data points within a data collection exhibit when compared to one another or the mean. The range, IQR, variance, and standard deviation are popular measurements of variability.

Four things make a problem statistical: the way in which you ask the question, the role and nature of the data, the particular ways in which you examine the data, and the types of interpretations you make from the investigation.

A statistics problem-solving process typically contains four components:

- Planning the problem (Ask a Question)
- Collect Data
- Analyse Data
- Interpret Results

All the activities in this chapter are built on this four-step method for resolving statistical issues. As you look at various statistical issues, this technique will become more and more familiar to you.

3. This is often referred to as starting the process with an expectation of variability. Investigations are successful when statistical questions are created that account for variability. For instance, each of the following statistical investigative questions foresees variability and can result in a thorough data collection process and subsequent data analysis:

- How fast can my plant grow?
- Do plants exposed to more sunlight grow faster?

- How does sunlight affect the growth of a plant?

For plants to survive, three things are essential: carbon dioxide, water, and sunlight. The plants use the energy from the sun to turn carbon dioxide, soil nutrients, and water into food through a process known as photosynthesis! In this project, we will:

1.watch as seeds sprout and monitor plant growth as 'Basil' (a seasoning herb) plants appear.

2.follow the development of basil seeds under three different lighting conditions—full sun, some sun, and limited/ no sun—and see how photosynthesis works! Think about these crucial questions before we start:

- Can a seed grow/germinate/develop into a plant with limited or no sunlight?
- How do you think a seed will grow with some or partial sunlight?
- What do you think the plant will look like after two weeks of growth?
- What will be the difference between the three sunlight exposure plants?
- How do you think the plants will be alike?

In contrast, the question 'How tall is the plant?' is answered with a single height, it is therefore not a statistical investigative question. 'How tall is the plant?' is a question we ask to collect data.

Higher Order Thinking Skills (HOTS)

1. Here are five statistical investigative questions to determine a student's immunity to catching cold and flu based on the given scenario:

   i. What is the overall percentage of students in the class who have been affected by cold and flu in a semester?

   ii. Is there a correlation between a student's age and their susceptibility to cold and flu?

   iii. Are there any identifiable patterns in the times of the year when students are more likely to catch cold and flu?

   iv. Does the frequency of catching cold and flu differ between students who live on campus versus those who commute?

   v. Is there any relationship between a student's sleep duration and their likelihood of contracting cold and flu?

2. Do it yourself.

3. Identifying Patterns

Unsolved Exercise

Objective Type Questions

A. 1. d 2. a 3. c 4. a 5. a

6. d 7. b 8. a 9. c 10. c

11. d 12. a 13. a 14. b

B. 1. parameter, statistic 2. n 3. Random sampling

4. Sample Frame Error 5. Probability

Standard Questions

A. 1. Facts are true and cannot be argued with, because they can be proven and are supported by evidence, while opinions vary according to the attitudes of the writer. But keep in mind that information might be manipulated to support the author's viewpoints. When people write articles, they often select facts that support their opinion.

2. Two types of prejudice are Racism and Ageism.

3. Partiality, preference and prejudice towards a set of data is called as a Bias.

4. An example of prejudice is having a negative attitude toward people who are not born in the United States. Bias is the tendency to favour one thought over another and maybe to ignore competing ideas.

5. The standard deviation is the variation in the population that is inferred from the variation in the sample.

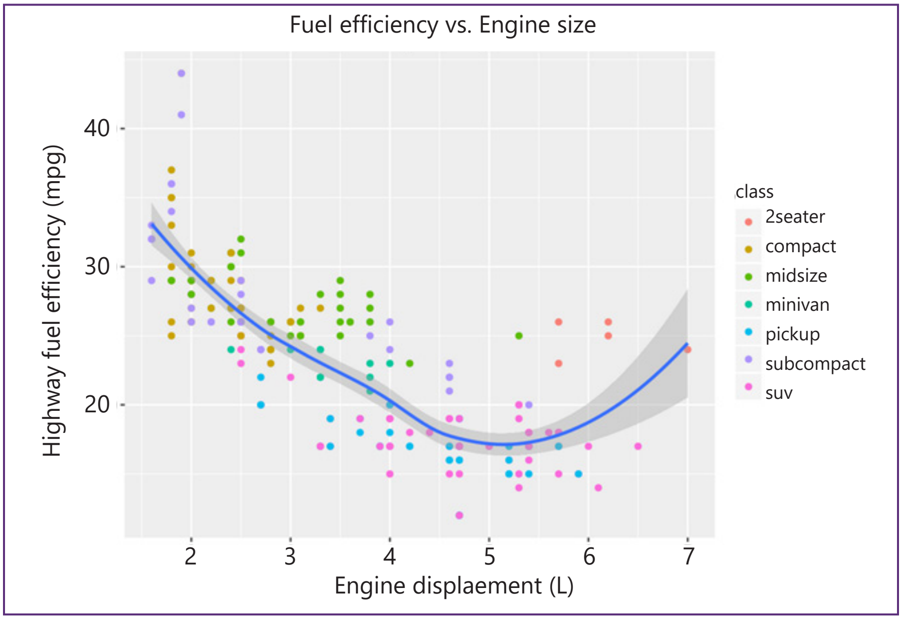
B. 1. According to the Central Limit Theorem, as sample sizes grow, the sampling distribution's form will always tend towards normalcy, regardless of how the population is distributed. This is helpful, as any research never knows which mean in the sampling distribution is the same as population mean, however, by selecting many random samples from population, the sample means will cluster together, allowing the researcher to make a good estimate of the population mean. Having said that, as the sample size grows, the error will always decrease.

2. Recall Bias is a type of measurement bias. It frequently occurs during the data labelling phase of any project. When you inconsistently classify comparable types of data, you have this form of bias. Thus, resulting in lower accuracy. For example, let us say we have a team labelling images of damaged laptops. The damaged laptops are tagged across labels as damaged, partially damaged, and undamaged. Now, if a team member marks an image as damaged and another one that is identical to it as somewhat damaged, your data will obviously be inconsistent.

3. Linearity Bias is the belief that changing one quantity would automatically result in a corresponding change in another. Unlike Selection Bias, Linearity Bias is a cognitive bias; it's produced not through some statistical process, but rather via how we mistakenly interpret the world around us.

   For example, let us take the case of relationship between fuel efficiency and engine displacement (i.e. engine size) of automobiles?

   Automobile engineering tells us that, as engines become larger, their fuel efficiency decreases. The majority of people believe that the relationship between engine displacement and fuel economy is a straight line. It is, at first, but real data tells a more complex story for the full picture, as shown in following figure:



   Putting all vehicle classes together, the trend is nearly linear for engine sizes under 4.5 litres, but then the relationship between fuel efficiency and engine size is (unexpectedly) nonlinear.

4. Confirmation bias is something which does not happen due to the lack of data availability. It is a phenomenon wherein data scientists or analysts tend to lean towards data that is in alignment with their beliefs, views, and opinions.

   They often focus knowledge from facts that expedites their proposal or hypothesis while filtering information; the moment they discover information that even marginally refutes their speculation, they turn away from it.

   Information that doesn't meet a data scientist's predefined view must be discarded. It is important to take in new data with an open mind. This phenomenon is progressively normal among authoritative organisations who want to assign importance to their own perceptions. Confirmation bias frequently results in poor business outcomes, which is the reason you should pay special attention to non-confirming proof.

5. The most used formula is:

$\mu_{\bar{x}} = \mu$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

6. Bias basically occurs because of sampling and estimation. If we had complete knowledge of every entity in our database and had information on every potential entity, our data would never have any bias. However, data science is often not conducted in carefully controlled conditions. It is mostly done of the "found data", i.e., the information gathered without regard to modelling. That is the reason why this data is very likely to have biases.
7. Z is the Z-score corresponding to the desired confidence level. For a 99% confidence level, $Z \approx 2.576$ (from the standard normal distribution).

$$\begin{aligned}\text{Sampling Error} &= Z \times \frac{\sigma}{\sqrt{n}}\\ &= 2.576 \times \frac{0.2}{\sqrt{36}}\\ &= 2.576 \times \frac{0.2}{6}\\ &= 2.576 \times 0.0333\\ &\approx 0.0858\end{aligned}$$

8. Do it yourself.

Higher Order Thinking Skills (HOTS)

1. The main cause of this type of bias is skin colour. It is categorized into selection bias.
2. Do it yourself.

4. Data Merging

Unsolved Exercise

Objective Type Questions

A. 1. a 2. c 3. a 4. c 5. a

B. 1. One to One 2. first quartile 3. 0, 1

4. 75th 5. 10

Standard Questions

A. 1. One to Many join is used to create a join between two tables. Any single row of the first table can be linked to one or more rows of the second table, but the rows of the second table can only link to the only row in the first table. It is also known as a many to one join.

2. The z-score is important because it tells you not only about the value but also where it falls in the distribution. It is very useful to standardise the values of a normal distribution by changing them into a z-score because:

   - It gives us a chance to calculate the probability of a value occurring within a normal distribution.
   - Z-score permits us to compare two values that are from different samples.

3. Quartiles are a type of percentile. They are a set of descriptive statistics. They summarise the central tendency and variability of a dataset or distribution.

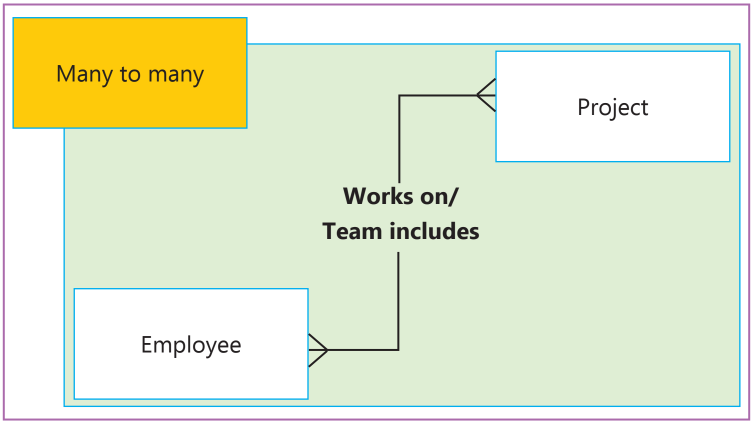
4. The second quartile (Q2, or the median) is the 50th percentile, which means that 50% of the data falls below the second quartile.

5. Decile is a technique that is used to divide a distribution into ten equivalent parts. When data is divided into deciles, a decile rank is allotted to each data point in order to sort the data into an ascending or descending order.

B. 1. Many to Many join is many to many joins that create a link between two tables. Each row of the first table can link to any row (or no row) in the second table. In the same way, each row of the second table can also link to more than one row of the first table. It is also represented as an N:N relationship.

In this join, none of the associated tables have primary key column. It means that all the columns of primary key fields are associated with all the columns of related tables. It occurs when each record in Table A may have many linked records in Table B and vice-versa.

For example, there are many people associated with each project, and every person can associate with more than one project.



2. By using the values of the quartiles, we can also calculate the interquartile range. An interquartile range is defined as the measure of the middle 50% of the values when ordered from lowest to highest. The interquartile range can be determined by subtracting the first quartile (Q1) from the third quartile (Q3).

   IQR = Q3 – Q1

   Where:

   IQR = interquartile range

   Q1 = 1st quartile or 25th percentile

   Q3 = 3rd quartile or 75th percentile

3. The value of the z-score states you how many standard deviations you are away from the mean.
   - If a z-score is equivalent to 0, it is on the mean.
   - A positive z-score specifies that the raw score is higher than the mean average. For example, if a z-score is equal to +3, it is 3 standard deviations above the mean.
   - A negative z-score tells that the raw score is below the mean average. For example, if a z-score is equal to –5, it is 5 standard deviations below the mean.

4. To calculate the z-score, you can use the formula:

   $$Z = \frac{(x - \mu)}{\sigma}$$

   Plugging in the values:

   $$Z = \frac{(105.5 - 122)}{6.2}$$

   $$Z = \frac{-16.5}{6.2}$$

   $$Z \approx -2.66$$

So, the z-score of the value 105.5 in the given dataset is approximately -2.66.

5. The first ten prime numbers are: 2, 3, 5, 7, 11, 13, 17, 19, 23, 29.

1. Q1 position = $\frac{(10 + 1)}{4} = \frac{11}{4} \approx 2.75$

- Since the position is not a whole number, round it up to the nearest whole number. So, Q1 is the value at the 3rd position, which is 5.

2. Q3 position = $\frac{3(10 + 1)}{4} = \frac{33}{4} \approx 8.25$

- Again, since the position is not a whole number, round it up to the nearest whole number. So, Q3 is the value at the 9th position, which is 23.

Now, calculate the interquartile range (IQR):

$IQR = Q3 - Q1 = 23 - 5 = 18$

So, the interquartile range for the first ten prime numbers is 18

Higher Order Thinking Skills (HOTS)

1. To compare the student's performance on the both exams, we can calculate the z-score for each test score. The z-score tells us how many standard deviations a data pint is form the mean.

for the first test:

- Test score: 85
- Mean: 60
- Standard deviation: 15

$$Z_1 = \frac{(X_1 - \mu_1)}{\sigma_1} = \frac{(85-60)}{15} = \frac{25}{15} = \frac{5}{3}$$

For the second test:

- Test score: 80
- Mean: 54
- Standard deviation: 12

$$Z_2 = \frac{(X_2 - \mu_2)}{\sigma_2} = \frac{(80-54)}{12} = \frac{26}{12} = \frac{13}{3}$$

Now, let's compare the z-score:

- $Z_1 = \frac{5}{3} \approx 1.67$
- $Z_2 = \frac{13}{6} \approx 2.17$

Comparing the z-scores, we see that the z-score for the second test (2.17) is higher than the z-score for the first test (1.67). This indicates that the student's performance relative to their peers was better in the second test compared to the first test.

In conclusion, the student performed relatively better in the second test compared to the first test, based on their z-scores.

Applied Project

Do it yourself.

5. Ethics in Data Science

Unsolved Exercise

Objective Type Questions

A. 1. a 2. a 3. b 4. d 5. b 6. b

B. 1. True 2. True 3. True 4. False 5. False

Standard Questions

A. 1. Techniques of safely discarding digital confidential data are Hard Drive and Tape Shredding, Hard Drive Degaussing, and Hard Drive Erasure.

2. Techniques of safely discarding physical confidential data are
   - Use scissors or a hammer to destroy the chip embedded in the Plastic-based Records (Debit /Credit Cards)
   - Shredding the Paper-based Documents
   - Cutting up the Documents
   - Burning the Documents
   - Incineration

3. Once you are done with the user data, especially confidential data, it is important that you discard this data in appropriate way to make sure that it is not accessed by any unauthorised person and it is not misused in anyway.

4. The most secure and cost efficient method is shredding, to dispose of all types of end-of-life hard drives and media tapes. The hard drive shredding services is great for businesses with large data centres or a stockpile of old hard drives and media tapes as it is a quick and effective technique.

5. Some examples of PII include:
   - Full name
   - Birthdate
   - Street address

B. 1. Note that in most of the devices, if you do a soft delete of a particular file, this file removes data from the original place and gets stored in a temporary folder from where one can easily restore these files. Hence, it is important that confidential data is cleaned out or formatted from the disk permanently and nobody can recover the files we destroyed.

2. Shredding is an excellent technique to erase data if you have a huge enterprise data centre or a large stockpile of old media that you want to destroy. It's very secure, fast and efficient. Shredding reduces electronic devices to pieces no larger than 2 millimeters. If you operate with highly secure data in a highly secure environment, shredding should be your number one choice as it guarantees that all data is obliterated. In circumstances when you just have one page or one file to discard, cutting the documents can be an appropriate method to discard the documents. While cutting the documents, you should cut them into small pieces and ensure that none of the sensitive information is readable. Also, you should cut the document in a way that it is not in a position to be rebuilt and it is completely unreadable. If these conditions are met, you have successfully discarded the data.

3. Burning the papers is likewise regarded as a reliable method of getting rid of documents as it makes sure that the documents that are burnt can never be reconstructed or read again. Although this method is not always practical, it is beneficial many a times when no other means of discarding are conveniently available. In terms of pollution, this approach is not advised.

   Incineration is a method of treating waste which involves burning the organic compounds found in waste products. The same can be used for paper and plastic-based materials. The solid mass of the original waste is reduced by around 80 to 85%, while the volume is reduced by around 95 and 96%. During burning of waste, however, pollutants are created. However, there are other pollution-free incinerators available.

4. There are at least nine Data Governance strategic objectives:

   i. Create an information-centric and informed organisational culture.

   ii. Establish a data governance program to provide accountability for information assets.

   iii. Provide for effective and appropriate information security.

   iv. Improve the quality and usefulness of information by making it timelier, more accurate, more complete and more accessible.

   v. Reduce the costs of managing information.

   vi. Share data through reusable processes; reuse data through shared processes.

   vii. Provide self-service business intelligence capabilities.

viii. Develop enterprise-class data management staff.

ix. Adopt enterprise-class data management tools

5. When managing data confidentiality, follow these guidelines:
   - Encrypt sensitive files
   - Manage data access
   - Physically secure devices and paper documents
   - Securely dispose of data, devices, and paper records
   - Manage data acquisition
   - Manage data utilisation
   - Manage devices

Higher Order Thinking Skills (HOTS)

1. Do it yourself.
2. Do it yourself.