K. K. Wagh Institute of Engineering Education and Research, Nashik.

Department of Computer Engineering

Academic Year: 2020 – 2021 Semester: I
Class & Div: BE – A and B
Course Name & Code: Laboratory Practice - I (410246)
Teaching Scheme: Practical (04 Hrs / week)
Name of Faculty: Prof. J. R. Mankar, Prof. A. V. Taware and Prof. N. G. Sharma

ASSIGNMENT 01 (DATA ANALYTICS) SAMPLE ORAL QUESTIONS

1. What is the type of dataset that is used as input to the program?
   Iris Dataset – csv file

2. How many attributes / features are there in the dataset?
   5 attributes :
   1.sepal length in cm
   2. sepal width in cm
   3. petal length in cm
   4. petal width in cm
   5. class:
   -- Iris Setosa
   -- Iris Versicolour
   -- Iris Virginica

3. What is the type of each attribute?
   All integer except class(species) which is categorical

4. How many features is numeric?
   4

5. How many features are nominal / categorical?
   1-species

6. Which function is used to display summary statistics in R?
7. What is difference between mean, median and mode?

Sl. No. -	Mean	Median	Mode
1.	The average was taken for a set of numbers is called a mean.	The middle value in the data set is called Median.	The number that occurs the most in a given list of numbers is called a mode.
2.	Add all of the numbers together and divide this sum of all numbers by a total number of numbers.	Place all the given numbers in an ascending order	It shows the frequency of occurrence.
3.	The result is the mean or average score.	The next step is to find the middle number on the list. It is called as the median.	We can have more than one mode or no mode at all.
4.	Example: To find the average of the four numbers 2, 4, 6, 8, we need to add the number first. 1. 2 + 4 + 6+ 8 = 20 2. Divide the sum by the total number of numbers, i. e 4. 3. 20/4 = 5 is the average or mean	Example: If the given list is 4, 2, 8, 10, 19. 1. Arrange the numbers in ascending order i .e 2, 4, 8, 10, 19. 2. As the total numbers are 5, so the middle number 8 is the median here.	Example: In the given series 3,3,5,6,7,7,8,1,1,1,4,5,6 1. Find the frequency of each number. 2. For number 3 it's 2, for 5 it's 2, for 6 it's 2, for 7 it's 2, for 8 it's one, for 1 it's 3, for 4 it's 1. 3. The number with the highest frequency is the mode.

8. What is meant by first, second and third quartile?

   The first quartile (Q_1) is defined as the middle number between the smallest number (minimum) and the median of the data set. It is also known as the *lower* or *25th empirical* quartile, as 25% of the data is below this point.

   The second quartile (Q_2) is the median of a data set; thus 50% of the data lies below this point.

   The third quartile (Q_3) is the middle value between the median and the highest value (maximum) of the data set. It is known as the *upper* or *75th empirical* quartile, as 75% of the data lies below this point.

9. Which function is used to display histogram in R?
10. Which function is used to display box plot?
    1) df.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False, figsize = (8,8),notch=False)
    plt.show()

    2) plt.boxplot(df['petal_length'],notch=True)
    plt.show()

11. What is meant by range?
    it is **the difference between the highest and the lowest values in a set.**

12. What is meant by standard deviation?
    In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range.

13. What is meant by variance?
    Variance measures how far a data set is spread out. It is mathematically defined as **the average of the squared differences from the mean**.

14. What is meant by percentiles?
    A percentile (or a centile) is a measure used in statistics indicating the value *below which* a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found.

15. Are there any outliers in dataset?
    Yes in sepal width

16. What is an outlier?
    The extreme values in the data are called outliers. The outliers are a part of the group but are far away from the other members of the group.

17. When and why do we use histogram?
    Histograms are commonly used in **statistics to demonstrate how many of a certain type of variable occurs within a specific range**.

Use histograms when you have continuous measurements and want to understand the distribution of values and look for outliers. These graphs take your continuous measurements and place them into ranges of values known as bins.

18. When and why do we use box plot?
    A Box Plot is the visual representation of the statistical five number summary of a given data set.

    A Five Number Summary includes:

    Minimum
    First Quartile
    Median (Second Quartile)
    Third Quartile
    Maximum

19. Which function is used to display summary statistics in Python?
    -df.describe()

20. Which function is used to display histogram and box plot in Python?
    plt.hist(df['petal_width'],bins=20,color=['orange'])
    plt.xlabel("petal Width");
    plt.ylabel('Frequency')
    plt.show();

    df.plot(kind='hist', subplots=True,sharex=False, sharey=False, figsize = (8,8))

    for box plot refer q10.

Prof. J. R. Mankar, Prof. A. V. Taware and Prof. N. G. Sharma
Course Teacher