CE5033 Statistical Methods and Data Mining

20240227 Exercise

1. What is the primary purpose of R?

   (A) Web development (B) Statistical computing and graphics

   (C) Mobile app development (D) Game development

   Reason: R is primarily designed for statistical analysis and graphical representation of data. It is not focused on web, mobile app, or game development.

2. Which of the following is a feature of R?

   (A) Limited statistical techniques (B) No support for graphical techniques

   (C) Comprehensive suite of statistical and graphical techniques

   (D) Closed source development

   Reason: R is known for its wide range of statistical and graphical capabilities, supporting various types of analyses and visualizations. It is not limited in statistical techniques, does support graphical techniques, and is open-source.

3. Why is R considered important in data science?

   (A) It has a limited package repository (B) It lacks community support

   (C) For its extensive statistical and visualization capabilities (D) It is not open-source

   Reason: R's wide array of packages and community contributions make it a powerful tool for data analysis, including advanced statistical methods and data visualization, which are crucial in data science.

4. What is a variable in R?

   (A) A constant value that cannot change (B) A function to perform operations

   (C) A package for statistical analysis

   (D) A storage place for data that can be changed or reused

   Reason: In programming, a variable is used to store data that can be modified or accessed later, and R follows this definition.

5. What is the advantage of vectorized operations in R?

   (A) They require explicit loops to perform operations.

   (B) They are slower but easier to understand.

   (C) They perform operations element-wise on vectors without the need for loops.

   (D) They can only be used with numeric data types.

Reason: Vectorized operations in R are designed to operate on whole vectors at once, making code more concise and often faster than equivalent code using explicit loops.

6. How do you create a vector in R?

   (A) vector = c(1, 2, 3) (B) vector = [1, 2, 3] (C) vector = {1, 2, 3} (D) vector = <1, 2, 3>

   Reason: In R, the "c()" function is used to combine values into a vector, making option A the correct syntax.

7. How do you access the 3rd element of a vector named "vec"?

   (A) vec(3) (B) vec[3] (C) vec{3} (D) vec<3>

   Reason: In R, elements of a vector are accessed using square brackets [] with the index of the desired element.

8. What is the correct way to create a matrix in R?

   (A) matrix = [1, 2; 3, 4] (B) matrix = matrix(c(1, 2, 3, 4), nrow = 2)

   (C) matrix = {1, 2, 3, 4} (D) matrix = <1, 2, 3, 4>

   Reason: The matrix() function in R takes a vector of values and the desired number of rows (or columns) to create a matrix, making option B the correct syntax.

9. What is the primary purpose of Exploratory Data Analysis (EDA)?

   (A) To confirm the hypothesis. (B) To understand the dataset's structure and characteristics.

   (C) To create predictive models. (D) To publish research findings.

   Reason: EDA is a crucial step in data analysis to explore and summarize the main characteristics of the dataset, often visually, before applying more formal statistical analysis.

10. Which data structure allows you to store elements of different types?

    (A) Vector (B) Matrix (C) Data frame (D) Factor

    Reason: Data frames in R can hold columns of different data types, unlike vectors and matrices which require elements of the same type.

11. How do you subset the first column of a data frame named "df"?

    (A) df[1] (B) df[,1] (C) df[1,] (D) df{"1"}

    Reason: In R, data frame columns can be accessed using the "df[, column_index]" notation, where "column_index" is the index of the column.

12. Which of the following is true about factors in R?

    (A) Factors are used for numeric data only. (B) Factors can only have two levels.

    (C) Factors are used to represent categorical data. (D) Factors cannot be ordered.

    Reason: Factors in R are used to encode categorical variables with a fixed number of levels, and they can be ordered or unordered.

13. Which of the following is true about the median?

(A) It is the most frequent value in a dataset.

(B) It is the middle value when the data set is ordered.

(C) It is the sum of all values divided by the number of values.

(D) It is affected significantly by extreme values.

Reason: The median is defined as the middle value in an ordered dataset, ensuring that half of the data points are below it and half are above it. This characteristic makes it particularly useful in understanding the central tendency of a dataset that may not be symmetrically distributed. Unlike the mean, the median is not influenced significantly by extreme values (outliers), making it a robust measure of central tendency for skewed distributions

14. Which figure type is most appropriate for visualizing the distribution of a single continuous variable?

(A) Line chart (B) Pie chart (C) Scatter plot (D) Histogram

Reason: Histograms are used to visualize the distribution of a continuous variable by showing the frequency of data points within specified ranges or bins.

15. When you want to show the relationship between two continuous variables, what is the best type of figure to use?

(A) Scatter plot (B) Histogram (C) Pie chart (D) Bar chart

Reason: Scatter plots are ideal for displaying the relationship between two continuous variables, where each point represents an observation with two measurements.

16. For comparing the mean values of a continuous variable across different categories, which figure would you use?

(A) Histogram (B) Scatter plot (C) Bar chart (D) Box plot

Reason: Bar charts are useful for comparing the mean (or other summary statistics) of a continuous variable across different categorical groups.

17. For exploring the distribution and outliers of a continuous variable, which figure is most informative?

(A) Bar chart (B) Histogram (C) Pie chart (D) Box plot

Reason: Box plots (or box-and-whisker plots) are designed to show the distribution of a continuous variable, highlighting the median, quartiles, and outliers.

18. To show the change in unemployment rate over time, the best type of figure to use would be a:

(A) Pie chart (B) Bar chart (C) Line graph (D) Scatter plot

Reason: Line graphs are excellent for illustrating trends over time, such as changes in the unemployment rate, by connecting individual data points with lines.

19. What measure of central tendency is most affected by outliers?

(A) Mean (B) Median (C) Mode (D) Range

Reason: The mean (or average) is sensitive to outliers because it is calculated by summing all values and dividing by the number of observations, making it susceptible to extreme values.

20. Which of the following is a measure of variability?

(A) Mean (B) Median (C) Standard Deviation (D) Mode

Reason: The standard deviation is a measure of the amount of variation or dispersion in a set of values, indicating how spread out the data points are from the mean.