Summer 2024 TTPR Bootcamp

Preassessment - Knowledge

For this preassessment, please do not use any outside resources or references. Answer them as best as you can. It is perfectly ok to also leave questions blank if you don't think you can answer them.

Answer the following questions as best as you can. For any open ended questions, you only need to give 2-4 sentences.

1. What is data science?

2. What are some reasons people use data science today?

3. What is data cleaning?

4. What are some reasons people perform data cleaning?

5. Match the statistical concept to the correct definition. Write the letter of the definition in the [] next to the concept.

| **Concept** |  | **Definition** |
| --- | --- | --- |
| [] Mean |  | a) Value that appears most often in a dataset. |
| [] Mode |  | b) Difference between the lowest and the highest value. |
| [] Range |  | c) Measurement of dispersion, calculated using the square root of the variance. |
| [] Median |  | d) Total of all values divided by the number of values. |
| [] Variance |  | e) Middle value in a list ordered from smallest to largest. |
| []Standard Deviation |  | f) Measurement of the spread of values in a dataset, calculated by the average of the squared differences from the mean. |

6. Find the median in the following list of numbers: 1, 3, 4, 5, 7, 8, 9, 12

7. Match the type of variable with the correct definition. Write the letter of the definition in the [] next to the variable name.

| **Variable Name** |  | **Definition** |
| --- | --- | --- |
| [] Categorical |  | a) Numerical variables that represent a measurement of quantity. |
| [] Ordinal |  | b) Variables with a numerical value and can be measured along a continuum. |
| [] Discrete |  | c) Interval variable that has a meaningful zero. |
| [] Continuous |  | d) Variables that have at least two categories without intrinsic order. |
| [] Binary |  | e) Numeric variables that take any value in an infinite range. |
| [] Ratio |  | g) Numeric variables that have a finite number of values. |
| [] Interval |  | h) A variable that can be put into categories, also known as a nominal variable. |
| [] Nominal |  | i) Variables that can be ranked. |
| [] Qualitative |  | j) Variable that is non-numerical with data that fits into categories. |
| [] Quantitative |  | f) Nominal variables that only have two categories or levels. Also referred to as dichotomous. |

Consider the following table:

| **Cafeteria Menu** | | | | |
| --- | --- | --- | --- | --- |
| **Entrée** | **Type** | **Total Calories** | **Protein (g)** | **Sugar (g)** |
| Turkey Sandwich | Cold | 400 | 25 | 5 |
| Spaghetti and Meatballs | Hot | 750 | 20 | 15 |
| Cesar Salad | Cold | 350 | 5 | 7 |
| Grilled Cheese | Hot | 625 | 15 | 10 |

8. What are the individuals in this dataset?

1. The Cafeteria’s customers
2. Entrée
3. Type
4. Menu

Your answer:

9. How many variables are there, and of those variables, how many are categorical?

1. 5 variables, 2 categorical
2. 5 variables, 1 categorical
3. 4 variables, 1 categorical
4. 4 variables, 0 categorical

Your answer:

What is the general use of each of the following Python libraries:

10. Scikit-learn

11. NumPy

12. SciPy

13. Matplotlib

14. Seaborn

15. Pandas

16. Match the definition and the concept. Write the letter of the definition in the []

[] Data normalization

[] Data formatting

[] Data binning

a. helps make variables comparable and helps eliminate inherent biases in statistical models

b. a technique to improve model accuracy and data visualization

c. critical for making data from various sources consistent and comparable

17. Match the graph type to its use. Write the letter of the definition in the []

[] Histograms

[] Box plots

[] Scatter plots

[] Heatmaps

a. exploring relationships between continuous variables, like engine size and price, in a car data set.

b. visual representation of the data's distribution for numerical data, indicating features like the median, quartiles, and outliers.

c. comprehensive visual summary of the strength and direction of correlations among multiple variables.

d. visualizing the distribution of binned data and gain insights into feature distributions

18. Name the *pandas* method used to group rows of a dataframe by one or more columns and then apply a function to the grouped data. This can be used to explore relationships between categorical variables.

19. Name the *pandas* method used to reshape a dataframe by pivoting one or more columns. This can be used along with heat maps for better data visualizations.

20. Correlation between variables is:

1. a statistical measure that indicates how the changes in one variable cause changes in another variable.
2. a statistical measure that indicates how the changes in one variable might be associated with changes in another variable.
3. a statistical measure that indicates how one variable provides information about another variable.

Your answer:

21. What is Pearson correlation?

22. What does a correlation coefficient close to 1, -1, and 0 indicate?

23. What is the difference between a train set and a test set?

24. How does cross-validation work?

25. Match the definition. Write the letter of the definition in the []

[] Overfitting

[] Generalization error

[] Underfitting

a. focuses too much on training data, including outliers and noise

b. unable to accurately capture the relationship between input and output variables in a dataset, generating a c. high error rate on both the training set and unseen data

d. measures how well your data does at predicting previously unseen data

26. Name the machine learning technique that predicts discrete labels or categories

27. Name the machine learning technique that predicts continuous numerical values

28. Name the machine learning technique that groups unlabeled data in groups that are similar in terms of the input attributes

27. Give causes and/or examples of machine learning bias

28. Name example applications of linear regression, classification, and/or clustering?