EECS16ML: Quiz

Note: Please explain in detail for all answers for credit.

1. Consider the problem of finding the sum of a certain column value in a dataset. Which of the following method would be best in this case:

A) Simple For Loop
B) List Comprehension
C) Pandas Aggregation

2. Consider a function that cleans up the column names of a dataset. Which of the following options follows proper naming and documenting practices:

A) CleanColumnNames
B) clean_column_names
C) Clean_Columns

3. Consider you are working with a dataset and are tasked with creating a simple reusable block of code that turns a column with categorical variables to one with numerical values that regression models can be applied on. How would you do this efficiently?

4. Assume you were given a scatter plot with a slight correlation but one that is hard to see due to the scaling and most of the points are in the bottom right of the graph. Which of the following scalings would be appropriate for a better visualization: (Select all that apply)

A) X**2
B) np.log(X)
C) Y**2
D) np.log(Y)

5. When adding comments and documentation to your code, explain the main function of this practice?

6. Suppose we want to plot the proportions of people in a survey who chose one of 3 categories. Which of the following visualizations would be acceptable in this case?

A) Bar Plot
B) Box Plot
C) Heatmap

7. Suppose we want to visualize the percentiles, including the mean and quartiles, of scores on the SAT. Which of the following visualizations would be acceptable in this case?

A) Bar Plot
B) Box Plot
C) Heatmap

8. Suppose we want to visualize a matrix which contains information about the intensities of different combinations of chemicals. Which of the following methods would be the best way to do so?

A) Bar Plot
B) Box Plot
C) Heatmap

9. Suppose given a certain dirty dataset with missing values, you want to get rid of any columns that contain empty values. Which of the following methods would be the best way to do so?

A) df.dropna()
B) df.dropna(how='all')
C) df.dropna(axis=1)
D) df.dropna(axis=0)

10. Assume you are creating a reusable simple pipeline with pdpipe that extracts the zip code given the string address. Which method would be most suitable in this case?