Dashboard / My courses / 911 - Data Informatics / DI 501 Section 1 / General / Quiz #3

Started on	Monday, December 6, 2021, 9:40 AM
State	Finished
Completed on	Monday, December 6, 2021, 10:00 AM
Time taken	19 mins 59 secs
Grade	**6.67** out of 10.00 (**67**%)

Question **1**

Partially correct

0.50 points out of 1.00

Regarding a demographic dataset, please match each of the cases with the corresponding data quality dimension.

When a person hits the age of 18, his/her majority status is not updated.	Consistency	✗
A person has an age of 200.	Reasonableness	✓

The correct answer is:
When a person hits the age of 18, his/her majority status is not updated. → Currency,

A person has an age of 200. → Reasonableness

Question **2**

Correct

1.00 points out of 1.00

Suppose we have 1000 temperature measurements periodically made by the same sensor in the last 1000 minutes (1 measurement in every minute). Most of the measurements vary between -5 and 40 but there are some suspicious measurements.

Match the following **techniques** applied to **the most appropriate problem** that might have necessitated the use of such a technique.

Find the first and third quantiles, Q1 and Q3, respectively and remove measurements that are larger than Q3 + 1.5 * (Q3-Q1) or smaller than Q1 - 1.5 (Q3-Q1).	Outlier ✓
Identify the frequent measurements which are entered as the same value across different attributes' values and remove them.	Inlier ✓
Round all values to nearest integers.	Noise ✓
Calculate the mean (m) and standard deviation (s) and drop values that are less than m-3*s or larger than m+3*s.	Outlier ✓

The correct answer is:
Find the first and third quantiles, Q1 and Q3, respectively and remove measurements that are larger than Q3 + 1.5 * (Q3-Q1) or smaller than Q1 - 1.5 (Q3-Q1). → Outlier,

Identify the frequent measurements which are entered as the same value across different attributes' values and remove them. → Inlier,

Round all values to nearest integers. → Noise, Calculate the mean (m) and standard deviation (s) and drop values that are less than m-3*s or larger than m+3*s. → Outlier

Question **3**

Incorrect

0.00 points out of 1.00

Suppose we want to train a classification model by using a dataset that consists of several examples (i.e., rows) with several real-valued features (i.e., columns) and a column that indicates the target label class.

By using some descriptive statistics, we discovered that variances of the columns are very large, there are some missing values in most of the columns but only 5% of the values in the entire dataset are missing.

We also discovered that the target label classes are very imbalanced.

Which of the following might be **the most appropriate method** to **deal with missing values** and **minimize the bias due to missing values** without using important information?

○ a. Use regression imputation. Train a regression model for each column using complete cases with the remaining columns and replace the missing values with the values predicted by the regression equation.

○ b. Use mean imputation across the rows. That is, replace the missing values in each row by the arithmetic mean of the non-missing values in the same row.

○ c. Use feature deletion. That is, remove columns that contain one or more missing values.

● d. Use mean imputation across the columns. That is, replace the missing values in each column with the arithmetic mean of the non-missing values in the same column. ✗

○ e. Use listwise deletion. That is, remove rows that contain one or more missing values.

The correct answer is:
Use regression imputation. Train a regression model for each column using complete cases with the remaining columns and replace the missing values with the values predicted by the regression equation.

Question **4**

Partially correct

0.50 points out of 1.00

There are 100 records (rows) in a dataset and each record has four features (columns) W, X, Y, and Z. In this dataset, X, Y, and Z are numeric features and W is a categorical feature. The records in this dataset constitute a random sample of a population of which some descriptive statistics are known.

In this dataset,

- All rows having missing values in column X also have missing values in column Y and vice versa.
- The means of attributes X and Y in the population are very close to the means of non-missing X and Y values in the dataset, respectively.
- The missingness of Z is not related to the missingness of any other feature.
- Missingness of W is not related to the missingness of any other feature.
- Z values are missing for all records on certain W values.
- The relative frequency distribution of non-missing W values in the dataset is very close to the relative frequency distributions of the W attributes of the population.
- The mean of attribute Z in the population is much higher than the mean of non-missing Z values in the dataset.

According to the information provided, which of the following **conclusions cannot be drawn**?

(You can select more than one).

☐ a. Missingness on Z is MNAR

☑ b. Missingness on Y is MCAR ✗

☐ c. Missingness on X is MAR

☑ d. Missingness on Y **is not** ignorable. ✓

☐ e. Missingness on W is MCAR

The correct answers are:
Missingness on Y **is not** ignorable.,

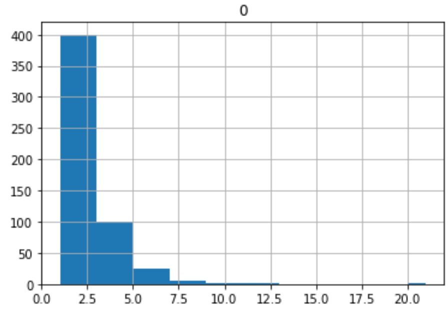
Missingness on X is MAR

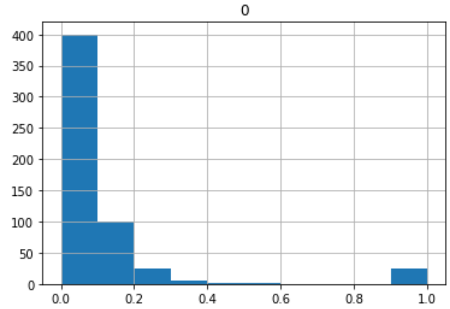
Question **5**

Correct

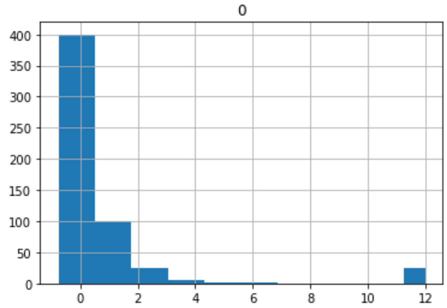
1.00 points out of 1.00

We have the histogram of the original data below. Please match the resulting histograms with the most probable technique applied to them.

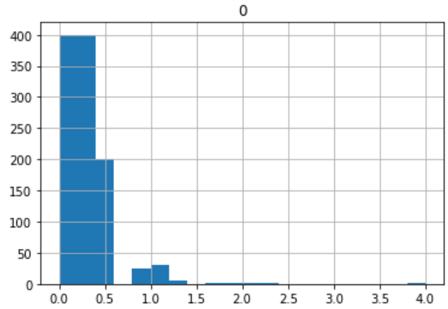




Normalization ✓

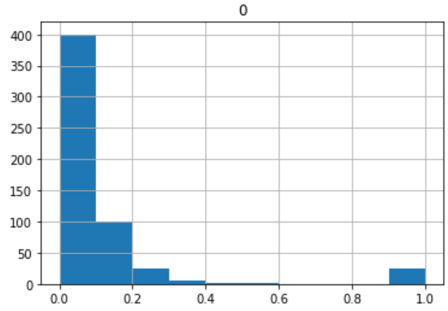


Standardization ✓

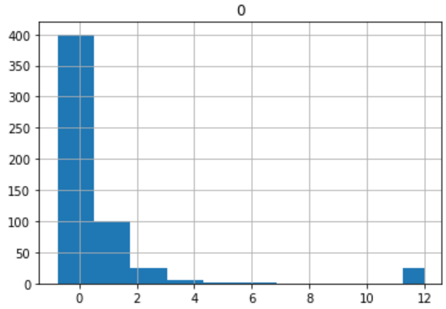


Rescaling (Adding (-1) and dividing by (5)) ✓

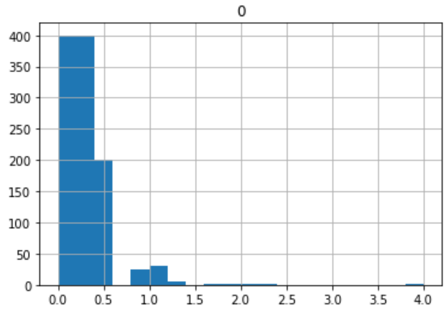
The correct answer is:



→ Normalization,



→ Standardization,



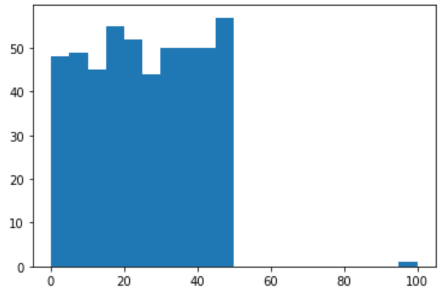
→ Rescaling (Adding (-1) and dividing by (5))

Question **6**

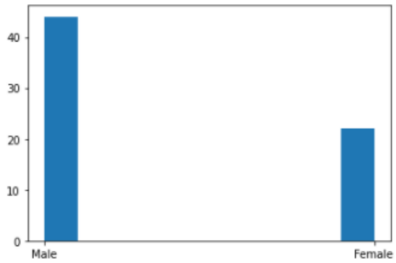
Partially correct

0.67 points out of 1.00

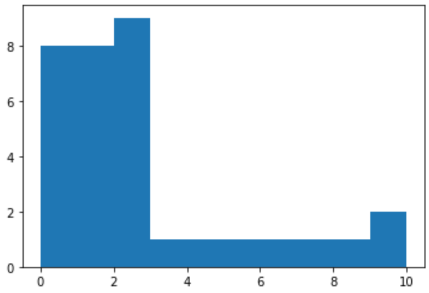
In the datasets whose histograms are given, please select whether using equal-width or equal-depth discretization is more likely to be used.



Equal-depth ✓

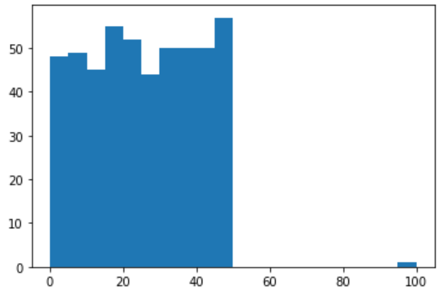


Equal-width ✓

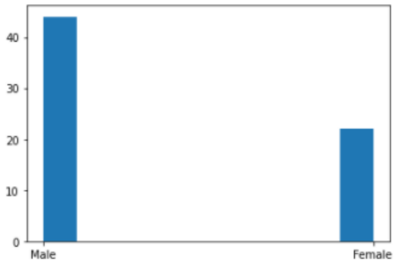


Equal-width ✗

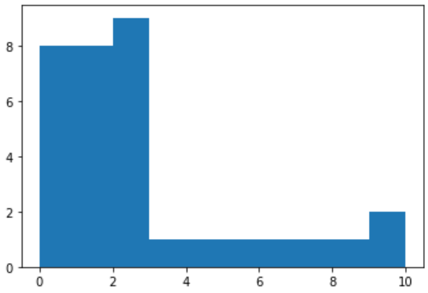
The correct answer is:



→ Equal-depth,



→ Equal-width,



→ Equal-depth

Question **7**

Correct

0.50 points out of 0.50

Outliers are generally harder to detect than inliers.

Select one:

○ True

● False ✓

The correct answer is 'False'.

Question **8**

Incorrect

0.00 points out of 0.50

One drawback of missingno's heatmap function is, if there isn't any missing values in an attribute/column, this attribute/column isn't shown in the heatmap.

Select one:

○ True

● False ✗

The correct answer is 'True'.

Question **9**

Correct

0.50 points out of 0.50

MICE imputation is generally done for datasets that have a low amount of missing values.

Select one:

○ True

● False ✓

The correct answer is 'False'.

Question **10**

Correct

0.50 points out of 0.50

Default ".duplicated()" function of pandas finds full duplicates rather than partial duplicates.

Select one:

● True ✓

○ False

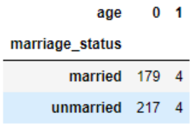
The correct answer is 'True'.

Question **11**

Partially correct

0.50 points out of 1.00

Regarding this crosstab output from the pandas library, which of those are correct? (You can select more than one).



age	0	1
marriage_status		
married	179	4
unmarried	217	4

☐ a. If those 0 and 1 is from missingness map, this table can be used for having an idea about missingness mechanisms of that variable.

☐ b. Even before applying a statistical test, as the distribution of marriage status looks unrelated to age group (or missingness), we can be somewhat sure they are unrelated.

☑ c. A contingency test may be used for checking the statistical importance of this table. ✓

☑ d. It says for age group 0 (or missingness), we have 179 married and 217 unmarried instances whereas, for age group 1 (or non-missingness), we have 4 married and 4 unmarried instances. ✓

The correct answers are:
If those 0 and 1 is from missingness map, this table can be used for having an idea about missingness mechanisms of that variable.,

A contingency test may be used for checking the statistical importance of this table.,

It says for age group 0 (or missingness), we have 179 married and 217 unmarried instances whereas, for age group 1 (or non-missingness), we have 4 married and 4 unmarried instances.,

Even before applying a statistical test, as the distribution of marriage status looks unrelated to age group (or missingness), we can be somewhat sure they are unrelated.

Question **12**

Correct

1.00 points out of 1.00

According to the different perspectives of data processing carried out using missingno and pandas packages in Python, which one of the below is correct?

● a. To find partial duplicates, we can use "subset" parameter in our .duplicated() function of pandas. ✓

○ b. Missingno's missingness matrix function does the sorting by null values inherently. We do not need to sort values to cluster null values together.

○ c. pd.to_numeric() function of pandas does not need additional paramater to correctly transofrm strings into numbers wherever possible and put null values whenever it is not possible.

○ d. ISO-XXXX-X encoding should be preferred over UTF-8 as it captures all of the different characters from different languages.

The correct answer is: To find partial duplicates, we can use "subset" parameter in our .duplicated() function of pandas.

◀ Quiz #2

Jump to...

Student Questions ▶