Quiz on data preprocessing

Due	Points	Questions	Available	Time Limit	Allowed Attempts
No due date	100	10	after Apr 20 at 12am	None	21

Take the Quiz Again

Attempt History

	Attempt	Time	Score
KEPT	Attempt 2	less than 1 minute	100 out of 100
LATEST	Attempt 2	less than 1 minute	100 out of 100
	Attempt 1	7 minutes	90 out of 100

Answers will be shown after your last attempt

Score for this attempt: **100** out of 100
Submitted Apr 21 at 9:31pm
This attempt took less than 1 minute.

Question 1

10 / 10 pts

Data inconsistency means data contains discrepancies in codes or names

- (•) True
- ( ) False

Question 2

10 / 10 pts

what problem we may face if the tuples with missing data are all deleted?

- (•) data may shrink dramatically
- ( ) it can cause inconsistency problem
- ( ) data may get noisier

Question 3

10 / 10 pts

Which data preprocessing task aims to handle missing data, smooth noisy data, identify or remove outliers, and resolve inconsistencies?

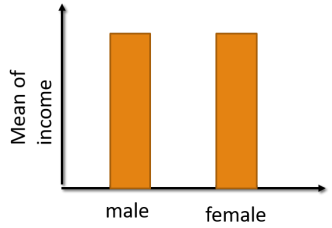
- ( ) data reduction
- (•) data cleaning
- ( ) data transformation
- ( ) data integration

Question 4

10 / 10 pts

Suppose we have a dataset with 100k tuples and 4 attributes:
age (integer, ranged from 18 to 90)
income (float, ranged from 0 to 1 billion)
job (categorical, 5 categories)
gender (binary, F/M).
We want to study if these attributes are related or independent by visualizing the data.

- gender VS income. (Is income related to gender?)



Is this a bar plot or histogram?

- histogram
- bar plot

Question 5

10 / 10 pts

Suppose we have a dataset with 100k tuples and 4 attributes:
age (integer, ranged from 18 to 90)
income (float, ranged from 0 to 1 billion)
job (categorical, 5 categories)
gender (binary, F/M).
We want to study if these attributes are related or independent by visualizing the data.

- gender VS income. (Is income related to gender?)

We visualized it by box plots. If the two box plots look the same for F and M, can we say they have the same distribution?

- True
- False

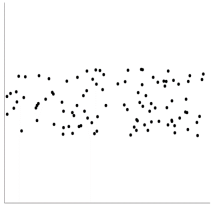
Question 6

10 / 10 pts

Suppose we have a dataset with 100k tuples and 4 attributes:
age (integer, ranged from 18 to 90)
income (float, ranged from 0 to 1 billion)
job (categorical, 5 categories)
gender (binary, F/M).
We want to study if these attributes are related or independent by visualizing the data.

- age VS income. (Is income related to age?)

We visualized it by scatter plot and get the following figure (x-axis is age. y-axis is income).



what conclusion is more likely to be correct?

- age and income are negatively correlated
- age and income are positively correlated
- age and income are not correlated

Question 7

10 / 10 pts

Suppose we have a dataset with 100k tuples and 4 attributes:
age (integer, ranged from 18 to 90)
income (float, ranged from 0 to 1 billion)
job (categorical, 5 categories)

gender (binary, F/M).
We want to study if these attributes are related or independent by visualizing the data.

- job VS income. (Is income distributed differently for different types of jobs?)

Between a bar chart and box plots, which plot do you think can give you more information on the income distribution?

- (●) box plot
- ( ) bar char

Question 8

10 / 10 pts

in the income attribute, we found there is a unit problem (some in US dollars, some in Euros). What category does this problem lie in?

- ( ) outlier
- (●) inconsistency
- ( ) incomplete

Question 9

10 / 10 pts

In box plot, the does the bar in the middle of the box mean?

- ( ) mean
- ( ) mode

- median

Question 10

10 / 10 pts

For data with n attributes, how many distinct scatterplots are there in the scatterplot matrix?

- n*n
- n(n-1)/2
- n(n-1)

Quiz Score: **100** out of 100