Data Analytics (CS40003)

Practice Set II
(Topic: Data Categorization)

I. Concept Questions

1. Classify the following attributes as library, discrete, or continuous. Also classify them as qualitative or quantitative. Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.
Example: Age in years
Answer: Discrete, quantitative, ratio.

(a) Time in terms of AM or PM. Dis I Qual
(b) Brightness as measured by a light meter. Con I Quan
(c) Brightness as measured by people judgments. Con I Quan
(d) Angles as measured in degrees between 0 to 360. Con I Qual
(e) Bronze, Silver and Gold medals as awarded at Olympics. cat
(f) Height above sea level.
(g) Number of patients in hospital. Dis I Quan
(h) ISBN numbers for books.
(i) Ability to pass light in terms of the following values: opaque, translucent, transparent.
(j) Military rank. Dis I Quan
(k) Distance from centre of IIT-KGP campus.
(l) Density of substance in grams per cubic centimetre.
(m) Coat check number.(when you attend an event, you can often give your coat to someone who, in turns, gives you a number that you can use to claim your coat when you leave.)

2. Mark as true or false the following statement. You should justify your answer with a suitable example in each answer.

(a) Range can be computed for any type of data. F. Not done for cat data
(b) Arithmetic mean of integer is always an integer. F.
(c) Probability of an impossible event is zero. true
(d) Regression is supervised learning model. true
(e) A regression through the origin indicates zero correlation among the variables. true
(f) Outliers can influence a regression model. true
(g) Every sample observation contributes to the regression coefficients. true
(h) Box plots are useless in providing strength of relationships. true

3. What type of data in NOIR typology are the following?

(a) Customer rating of a service with maximum score of 10. O
(b) Defaulting loan amount of a customer. R
(c) Jersey number of a player (to distinguish each player in a team). N
(d) Duration of a movie (in minute). R
(e) Answer choices for multiple-choice answers. N
(f) Week days coded as 0 = Sunday, 1 = Monday, …..,6 = Saturday. O
(g) Systolic blood pressure of a patient. R
(h) Resolution in dots per (dpi) of printers.
(i) Branches of studies in a collage. N
(j) Amount of sugar in a cup of orange juice. R
(k) Number of orders of an e-com customers R
(l) Wind classification as (breeze, gale, storm, and hurricane). Ord

4. Identify type of data, and assign single-letter codes for each of the following.

(a) hair color = {black, red, gray, white} . N
(b) credit card = {VISA, MasterCard, Bankcard}.N
(c) sales region = {urban, rural, village}. N
(d) auto insurance = {minimum liability, normal, premium, comprehensive}.or
(e) hobbies = {chess, music, movies, numismatics, philately, traveling}.N
(f) floppy = {3M, Dysan, IBM, Maxell, Sony, Verbatim}. N
(g) drink = {milk, fruit, juice, tea, coffee, cola}. N

5. What is the least informative of the scales?

(a) Nominal
(b) Ordinal
(c) Interval
(d) Ratio

6. On which type of data (N, O, I, R) can you compute each of following?

(a) Harmonic mean req +ve no. so R only
(b) Median of a sample O I R
(c) Range O I R
(d) Inter Quartile Range O I R
(e) Maximum of a sample O I R

7. What central tendency measures are most appropriate for nominal and ordinal data?

Nominal - mode
ordinal- mode & median

8. Identify each of the binary variables as symmetric or asymmetric:

(a) person = (left-handed, right-handed). as
(b) gender = (Male, Female). s
(c) diabetic = (Yes, No). as
(d) frequent Flyer = (Yes, No). as
(e) fever = (Yes, No). as

9. Identify each of the following variables as discrete or continuous.

(a) Formatted capacity of a computer disk. con
(b) Data transmission rate of an internet connection. con
(c) Number of words per minute typed by a clerk. con
(d) Foreign currency exchange rate on any day. con
(e) Total spam mail received per day. d
(f) Number of e-com orders from an IP address. d
(g) Total score accumulated by an online game player. d
(h) Calorific value of a soft drink. con
(i) Odometer reading of a car. con
(j) GDP of a country. con

10. What type of variable are the following?

(a) year = {regular year, leap year}. Nominal
(b) day = {work day, weekend, holiday}. Ordinal (not sure about the ordinal)
(c) entertainment = {TV, radio, music, movies}. Nominal
(d) blood pressure = {normal, abnormal}. Nominal
(e) travel time to school. ratio
(f) total spending on food per month. ratio
(g) Are the following the nominal variable?
(h) geographic position = {longitude, latitude, altitude}. Nominal
(i) ENT_consultation = {ear, nose, throat}. Nominal
(j) coffee = {water, coffee powder, milk, sugar}. Nominal
(k) tea type = {light, medium, strong}. ordinal

11. How you can convert data of interval type to ordinal type? Give an example. What are the issues with this transformation? Whether the reverse is possible? Justify your answer.

12. Give at least four points with which you can distinguish between interval and ratio data.

13. From the speed and storage point of views which type of data is/are most preferable? Give reasons to your answer.

II Objective Questions

1. Which cannot be measured with "Categorical" data?
- (a) Mean
- (b) Median
- (c) Mode
- (d) Variance

2. Consider the data about all students in a course stored with the following structure:

Name	Roll No	Category*	Mark1	Mark2	Total	Grade
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…

**Category denotes whether a student belongs to UG or PG*

If the structure is used to store the data of 100 students, then the dimension of the data is
- (a) 2
- (b) 7
- (c) 100
- (d) 200
- (e) 700

3. According to NOIR classification, the attribute "Category" in Table. Q2 can be categorized as
- (a) Categorical
- (b) Symmetric binary
- (c) Asymmetric binary
- (d) Ordinal

4. Which operation cannot be carried out on Ordinal data?
- (a) To find the minimum
- (b) To find the mean
- (c) To find the mode
- (d) To find the median

5. Which data scale used "zero point as origin"?
- (a) Nominal
- (b) Ordinal
- (c) Interval
- (d) Ratio

6. Which is not true?
- (a) Multiplication operation is not possible to interval data
- (b) Division operation is not possible to interval data
- (c) Only addition operation is possible for interval data
- (d) All arithmetic operations are applicable to interval data

7. Which is not true?

(a) Interval data can be transformed to "Categorical" data
(b) Interval data can be transformed to "Categorical" data and vice-versa
(c) Interval data can be transformed to ratio data
(d) Interval data can be transformed to ratio data and vice-versa

8. The data type that can be used to store both interval and ratio data are

(a) Integer
(b) Float
(c) Character
(d) Double

9. A data cube is used to model

(a) A multidimensional data
(b) Unstructured data
(c) Both multidimensional and multimodal data
(d) Only multimedia data

10. The streaming data from an earth satellite is

(a) Only multidimensional data
(b) Only multimedia data
(c) Both multidimensional and multimodal data
(d) Unstructured data