TDS2101

Data Science Fundamentals

Tutorial 2

1. Give some possible reasons for discretizing continuous-valued data by grouping them into “bins”.

2. Describe one possible weakness for each of the following methods that are used to handle missing values.

   (a) Ignore the tuple

   (b) Fill in missing value manually

   (c) Predict the missing values using learning algorithm

   (d) Use of a global constant to fill missing values

3. In the snippet of data shown in Table 1 below, identify the necessary step to be performed before the data is ready for constructing predictive models.

index	**Country**	**Land Area**	**GDP per capita**	**HD index**
1	Australia	7,633,565	51885	0.944
2	Canada	9,093,507	42080	0.929
3	Colombia	1,038,700	6432	0.767
4	Argentina	2,736,690	10006	0.845
5	Spain	498,980	29816	0.904

Table 1: World country statistics

4. Table 2 contains the selected list of suppliers that supply raw goods to the Nasi Lemak Enterprise.

State	Company-Name	Goods	Location	Postcode
Selangor	Ali & Friends Sdn Bhd	Eggs	Puchong	47100
	Ali & Friends Sdn Bhd	Peanuts	Puchong	47100
Kuala Lumpur	Alpha Store Enterprise	Anchovy, Banana leaf	Kampung Baru	50300
	Joe Speedmart Sdn Bhd	Rice	Pudu	55200

Melaka	Green Always Sdn Bhd	Banana Leaf	Alor Gajah	73000
Selangor	Spicy Farm Enterprise	Chilli	Bangi	43600
Selangor	Organic Food Sdn Bhd	Cucumber	Bangi	43000

Table 2: List of selected suppliers

(a) Convert Table 1 to 1NF

(b) Based on your answer in 4(a), convert the normalized table in 1NF to 2NF

(c) Based on your answer in 4(b), convert the tables in 2NF to 3NF

5. What are some advantages for storing data in a JSON file instead of a CSV file?

6. What do we gain from keeping data **immutable**?

7. Discuss the trade-off factors involved in storing your data in very raw form.