Faculty of Computers and Artificial Intelligence (Helwan University)

SW: Selected Topics in Software Engineering (Data Science and Big Data)

SHEET 1

1. "Validation and standardization are two strategies that are commonly used to smooth noisy data," explain briefly with the aid of examples the meaning of each of these terms.

2. Table 1 represents the collected data for a set of car models:

Table 1: Data that describe a set of car models

Car Model	Manufacturing Year	Cylinders	CC	Price (L.E.)	Maximum Speed (Km/hour)	Horse Power	Used/New	Usage Duration (Years)
Toyota	2009		1600	150000	195	1400	New	3
Jeep	2009	6	3700	320000	200	210	New	0
Mercury	MWVR09	6	4000	2500000		210		0
Opel	2008	4	1600	180000	192	105	New	10
Mitsubishi	2006	4	–1600		170	106	Used	3
Honda	2009	4	1500	120000	180	92	New	0
Mazda	2003	4	1600	85000	180		Used	6

a) Find any outliers in field "Price."
b) Fill-in the incomplete data in fields "Price" and "Maximum Speed" using mean/mode method.
c) Is there any inconsistency in the data values? If so, identify them and state the reasons behind these inconsistencies.

3. Table 2 shows the characteristics of a set of planets in the solar system.

Table 2: Characteristics of a set of planets in the solar system

Planet	Average Surface Temperature (°C)	Approximate Solar Day (Earth Days)	Approximate Solar Year (Earth Days)	Approximate Diameter (Km)	Average Distance to the Sun (Km)
Mercury	**179**	**58.65**	**88**	**4880**	**58000000**
Venus	**482**	**243**	**225**	**12102**	**108000000**
Mars	**–60**	**1**	**687**	**6792**	**230000000**
Saturn	**–153**	**0.4263**	**10760**	**12000**	**1427000000**
Uranus	**–218**	**0.7458**	**30681**	**50800**	**2870000000**
Pluto	**–222**	**6.3900**	**90545**	**2320**	**5900000000**

For the data shown in Table 2, use the following normalization methods to normalize "Approximate Solar Day", and "Approximate Diameter" fields:

i. Min-Max, use [0, 1] as the normalization interval
ii. Z-Score
iii. Decimal Scaling