**Sheet solution**

**Question 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**

Validation and standardization are ways to smoothing noisy data, it includes spelling check, e.g., kairo is incorrect and must be corrected to be Cairo. We can use dictionary on geographic names to solve such a problem.

Another example that we must validate the age by putting constraints like age must not be negative or zero, or greater the eighteen years. We must also validate ambiguous values, we can find that the abbreviation “s” refers to tow states “single” and “Separated”, which are ambiguous and must be validated.

Standardization is to have a uniform format, e.g., all words are in upper or lower case, remove prefixes and suffixes from names, Mohamed Ahmed instead of Mr. Mohamed Ahmed, Mohamed Ahmed instead of Mohamed Ahmed, Ph.D. Abbreviations and encoding schemes should consistently be resolved by consulting special dictionaries or applying predefined conversion rules, US is the standard abbreviation of United States.

**Question 2:**

1. Find any outliers in field “Price.”
   1. Q2(median)= 165,000
   2. Q1=120,000
   3. Q3=320,000
   4. IQR = Q3 - Q1=200,000
   5. lower limit = Q1 - 1.5\*IQR=-180,000
   6. upper limit = Q3 + 1.5\*IQR=620,000  
      **2,500,000 is outliers**
2. Fill-in the incomplete data in fields “Price” and “Maximum Speed” using mean/mode method
   1. Price: median  
      => 165,000
   2. Maximum Speed: median  
      => 186
3. Is there any inconsistency in the data values?
   1. In Manufacturing Year MWVR09 is not a year
   2. In CC -1600 must not be negative
   3. Toyota and Opel tuples have 3 and 10 usage years and have state new
   4. Toyota has 1400 horsepower which is too high
   5. Honda has 92 horsepower which is too low

**Question 3:**

1. Normalization
   * 1. Text

        Description automatically generated with medium confidence   
        A picture containing logo

        Description automatically generated

Logo

Description automatically generatedA picture containing shape

Description automatically generated

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

Diagram

Description automatically generated

A picture containing text, antenna

Description automatically generatedDiagram

Description automatically generated

* 1. Approximate Solar Day
     1. Min-Max
        1. minimum =0.4263(Approximate Solar Day)
        2. maximum =243(Approximate Solar Day)
        3. minimum =2,320(Approximate Diameter)
        4. maximum=50,800(Approximate Diameter)

|  |  |
| --- | --- |
| Approximate Solar Day (Earth Days) | Approximate Diameter |
| 0.240025 | 0.0528053 |
| 1 | 0.201774 |
| 0.00236505 | 0.0922442 |
| 0 | 0.19967 |
| 0.00131713 | 1 |
| 0.0245851 | 0 |

1. Z-Score

A̅ = 51.702016(Approximate Solar Day)

A̅= 14815.66(Approximate Diameter)

Standard deviation = 96.429(Approximate Solar Day)

Standard deviation = 18052.161(Approximate Diameter)

|  |  |
| --- | --- |
| **Approximate Solar Day (Earth Days)** | **Approximate Diameter** |
| 0.072053 | -0.817465 |
| 1.983821 | -0.807253 |
| -0.525796 | -0.820659 |
| -0.531746 | -0.820691 |
| -0.528432 | -0.820673 |
| -0.46990 | -0.82036 |

1. Decimal Scaling

Approximate Solar Day (Earth Days)/1000

Approximate Diameter/100000

|  |  |
| --- | --- |
| **Approximate Solar Day (Earth Days)** | **Approximate Diameter** |
| 0.05865 | 0.0488 |
| 0.243 | 0.12102 |
| 0.001 | 0.06792 |
| 0.0004263 | 0.12 |
| 0.0007458 | 0.508 |
| 0.00639 | 0.0232 |