Homework 1: Data

# Problem 1: Types of Attributes

Here we have a table with different types of data listed out. Label each one as either numerical or categorical. Then label it as either interval or ratio if it’s numerical or nominal or ordinal if it’s categorical. If it’s binary, label it as symmetric or asymmetric as well

|  |  |
| --- | --- |
| **Data Type** | **Examples** |
| Categorical, Ordinal | Rank in a race (1st place, 2nd place, 3rd place) |
| Numerical, Ratio | Weight in kilograms or pounds (70 kg, 150 lbs) |
| Binary, Symmetrical | Marital Status (single, or married) |
| Categorical, Nominal | Eye color (blue, brown, green) |
| Categorical, Nominal | Types of fruit (apple, banana, orange) |
| Numerical, interval | IQ scores (90, 110, 130) |
| Categorical, Nominal | Gender (male, female, non-binary) |
| Numerical, interval | Calendar dates (January 1, February 15) |
| Numerical, interval | Temperature in Celsius or Fahrenheit (20°C, 30°F) |
| Binary, Assymmetrical | Tested positive for breast cancer |
| Numerical, Ratio | Income in dollars ($50,000, $75,000) |
| Categorical, Ordinal | Educational attainment (high school, bachelor's, master's) |
| Categorical, Ordinal | Customer satisfaction rating (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied) |
| Categorical, Ordinal | Rank in a race (1st place, 2nd place, 3rd place) |
| Numerical, Ratio | Temperature in Kelvins |

# Problem 2

Normalize the following sets of numbers with Min-Max and Z-Score Normalization. These are the same examples from the slides in class. If you wrote down the answers, good job.

|  |  |
| --- | --- |
| Original Set | **12, 19, 23, 34, 41, 43, 56, 67, 78, 90**; mean = 46.3, sd = 26.0 |
| Min-Max | **0.00, 0.09, 0.14, 0.28, 0.37, 0.40, 0.56, 0.71, 0.85, 1.00** |
| Z-Score | **-1.32, -1.05, -0.90, -0.47, -0.20, -0.13, 0.37, 0.80, 1.22, 1.68** |
| Original Set | **5, 11, 18, 29, 32, 42, 47, 54, 63, 76**; mean = 37.7, sd = 23.0 |
| Min-Max | **0.00, 0.08, 0.18, 0.34, 0.38, 0.52, 0.59, 0.69, 0.82, 1.00** |
| Z-Score | **-1.42, -1.16, -0.86, -0.38, -0.25, 0.19, 0.40, 0.71, 1.10, 1.67** |
| Original Set | **16, 24, 28, 35, 40, 49, 61, 73, 87, 92**; mean = 50.5, sd = 26.7 |
| Min-Max | **0.00, 0.11, 0.16, 0.25, 0.32, 0.43, 0.59, 0.75, 0.93, 1.00** |
| Z-Score | **-1.29, -0.99, -0.84, -0.58, -0.39, -0.06, 0.39, 0.84, 1.37, 1.55** |
| Original Set | **14, 21, 27, 39, 45, 50, 59, 68, 72, 84**; mean = 47.9, sd = 23.1 |
| Min-Max | **0.00, 0.10, 0.19, 0.36, 0.44, 0.51, 0.64, 0.77, 0.83, 1.00** |
| Z-Score | **-1.47, -0.56, -0.44, -0.19, -0.06, 0.04 , 0.23, 0.42, 0.50, 0.75** |
| Original Set | **10, 17, 22, 31, 46, 58, 64, 75, 81, 93**; mean = 49.7, sd = 29.0 |
| Min-Max | **0.00, 0.08, 0.14, 0.25, 0.43, 0.58, 0.65, 0.78, 0.86, 1.00** |
| Z-Score | **-1.37, -1.07, -0.89, -0.58, -0.07, 0.35 , 0.56, 0.93, 1.14, 1.56** |

# Problem 3: Binarization, Discretization, Aggregation, and Sampling

This should be an easy one. Just match the term with the description like in Problem 1. For the ones where it’s sampling, list the type of sampling too

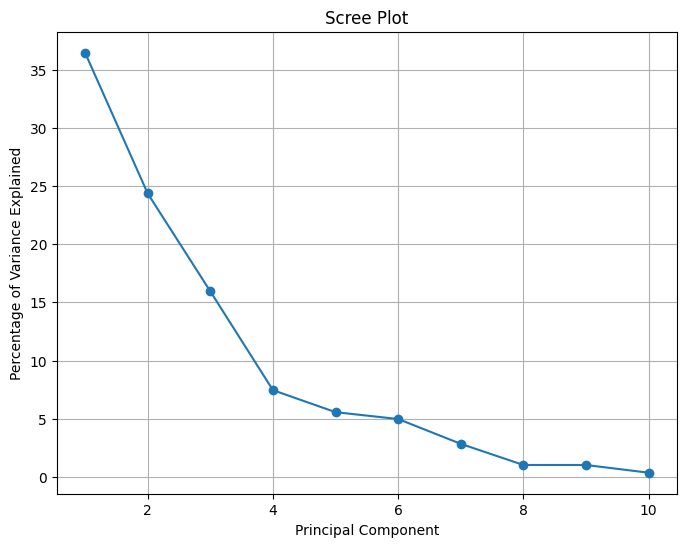
|  |  |
| --- | --- |
| **Binarization** | This method converts categorical data into asymmetric binary data |
| **Sampling: Stratified** | This utilizes partitioning and then drawing rows from each partition |
| **Aggregation** | This can reduce the data at the expense of insight |
| **Discretization** | This method takes continuous numeric data and groups it into discrete bins. |
| **Sampling: with replacement** | The chances of selecting a specific row are the same for every selection |
| **Aggregation** | This method combines multiple attributes or objects into one. |
| **Discretization** | This will be used when working with decision trees later in the semester |
| **Sampling: without replacement** | A specific row’s chances of being selected increase with each selection until it’s selected |
| **Aggregation** | This method can be used to change the granularity of the data |
| **Binarization** | This will be used when working with Market Basket Analysis |

# Problem 4: Principal Component Analysis

Here are some eigenvalues. Fill out the table

|  |  |  |
| --- | --- | --- |
| **Eigenvalue** | **Percentage of Variance Explained** | **Cumulative Percentage of Variance Explained** |
| 3.01012143 | 30.10 | 30.10 |
| 2.01123086 | 20.11 | 50.21 |
| 1.31674569 | 13.17 | 63.38 |
| 0.61562503 | 6.16 | 69.54 |
| 0.45954108 | 4.60 | 74.13 |
| 0.41082415 | 4.11 | 78.24 |
| 0.23463404 | 2.35 | 80.59 |
| 0.08481244 | 0.85 | 81.44 |
| 0.08479815 | 0.85 | 82.28 |
| 0.02991938 | 0.30 | 82.58 |

Here’s the corresponding scree plot. How many Principal Components should we use? Why?



|  |
| --- |
| Using 8 Principal Components. Because the line is getting flat after the eighth component, and it is only 1 percent increment at components number 9. |