CSE514, Fall 2022, HW 1 Name: Student ID:
Note: This homework is worth a total of 15 points

Q1 (2pts): Refer to the “Lunch features” dataset to give an example of each data type:

Q1a:	Numerical and Discrete:	Number of ingredients
Q1b:	Numerical and Continuous:	Price in dollars
Q1c:	Categorical and Nominal:	Culture of the food
Q1d:	Categorical and Ordinal:	Temperature (cold, warm, hot)

Q2 (5pts): You’ve been tasked with inputting the “Lunch features” dataset into a new database that can only accept numerical feature values. You must keep a minimum of 5 features in addition to price, but it’s fine to leave null values for samples that do not have a feature value recorded. List the features you’ll choose to keep and how you would process them for input:

Feature	Processing
Price	No processing necessary, just input the decimal value in dollar units
Q2a: Weight	No processing necessary, just input the numerical value in gram units
Q2b: Number of ingredients	No processing necessary, just input the numerical value in gram units
Q2c: Food culture	For each unique food culture, assign a number ID
Q2d: Cooking method	For each unique cooking method, assign a number ID
Q2e: Temperature	Define a range for temperature values (e.g. 0 = frozen and 10 = hot), and assign intermediates to maintain ordering

Q3 (2pts): Identify a data quality problem in the Spring subset of the “Lunch features” dataset. Propose a method to handle it.

Ex 1: Sample 26 (column Z in Excel) says “Hi”. It’s the only sample with this “feature,” and it’s unlikely to be useful for any relevant analyses of the dataset, so I would remove it while cleaning the dataset

Ex 2: The delivery method feature is spelled in multiple ways. I would define some replacement rules so that all samples can only have a value from a limited and defined set, such as {frozen, home-made, take-out, delivery, dine-in}

Q4 (6pts): Within the “Lunch features” dataset, the Spring subset has many more features than the Fall subset. To integrate the two into a single matrix, you could either drop all extra features from the Spring samples or add all the features to the Fall samples.
Answer three of the following with unique reasons:

Q4a: Why would dropping all extra features from the Spring samples would be a good idea?

We would reduce the sparsity of the dataset

Q4b: Why would dropping all extra features from the Spring samples would be a bad idea?

We would lose data that could be informative

Q4c: Why would adding all extra features to the Fall samples would be a good idea?

We would be able to include all the data from the Spring set in our analyses

Q4d: Why would adding all extra features to the Fall samples would be a bad idea?

We would increase the sparsity of the dataset