CSE514, Fall 2022, HW 1 Name: Ying (Fiona) Xu Student ID: 510576
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: Weight (eg. 300g)

Q1b: Numerical and Continuous: Price (eg. $7.02)

Q1c: Categorical and Nominal: Type (eg. American, Japanese)

Q1d: Categorical and Ordinal: Hot/Cold (eg. Hot, Cold)

Here I want to give the explanation of why wrote hot/cold: If all the values in this feature are hot/cold, we can see them as similar to high/medium/low.

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 (eg. 300)	No processing necessary, just input the integer value in gram units
Q2b: hot/cold (eg. hot)	Map them into the Boolean values, like hot = 1, cold = 0
Q2c: ingredients (eg. 3)	No processing necessary, just input the integer value
Q2d: take-out/ homemade/...	Firstly, give all kinds of features index, and put them into a list. And then use one-hot encoding to get the final weight matrix.
Q2e: type (eg. Asian)	Firstly, give all kinds of features index, and put them into a list. And then use one-hot encoding to get the final weight matrix.

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

Problem: The units of features are not uniform, for example, the weight of food (8 chicken nuggets VS 470g). When doing the data analyse, we should process the data in the same units, so that we can draw the correct conclusion.
Fixing Method: Before letting people input the "Lunch features" dataset, explain clearly the input feature, and give them the units and examples. For example, when inputting the weight of food, you should write as follows: This is a feature to tell the weight of food (You should input the integer value in gram units, eg. 300).

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?

Because the data in extra features from the Spring samples are sparse matrix. If we integrate the two into a single matrix, we will lose more data, and make the matrix sparser. Generally, we can not lead to some conclusion by using sparse matrix.

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

Because the data from the Fall samples are not enough. If we drop all extra features from the Spring samples, there will be less data for us to analyze.

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

Because the data from the Fall samples are less than those from the Spring samples, there will be fewer dimensions to analyze the data, so may draw a one-sided conclusion.

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

Because when merging two datasets, the data from the Fall samples lose some of the features compared with those from the Spring samples. Then, when analyzing the new merging matrix, a lot of data there will be None, which will cause some bad influencing from getting the conclusion.

Extra Credit: Calculate the partial derivative of parameter b

Extra Credit:

MSE loss function: $L(m,b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (mx_i+b))^2$

$$\frac{\partial L}{\partial b} = \frac{1}{n}\sum_{i=1}^{n} 2(y_i - (mx_i+b)) \cdot (0-(0+1))$$

$$= -\frac{1}{n}\sum_{i=1}^{n} 2(y_i - (mx_i+b))$$