Exercise 4.1 Part 1 Solution

Part 1. Concepts

Q4.1: Records Data format is the most common data format (but not the only) for Analytics work. Why?

The number and type of items in each column will be constant.
Each row represents unit of analysis.

Q4.2: What is the usefulness of transactions data format? How is this different from Records data format?

List all the items purchased or processed per transaction.
The number and type of items in each column will vary.

Q4.3: What is the difference between Nominal vs Ordinal? Why is this distinction important?

Nominal variable is a categorical variable without any order. The numbers serve only as labels or tags for identifying and classifying objects. The numbers do not reflect the amount of the characteristic possessed by the objects. Only a limited number of statistics, all of which are based on frequency counts, are permissible, e.g., percentages, and mode.

Ordinal variable is a categorical variable with order. It is a ranking scale in which numbers are assigned to objects to indicate the relative extent to which the objects possess some characteristic. Thus, we can determine whether an object has more or less of a characteristic than some other object, but not how much more or less. Any series of numbers can be assigned that preserves the ordered relationships between the objects. In addition to the counting operation allowed for nominal scale data, ordinal scales permit the use of statistics based on centiles, e.g., percentile, quartile, median.

Q4.4: Integers (aka whole numbers) may be treated as either Continuous or Categorical, depending on purpose. Provide an example for each situation. [Explain in your own words.]

5-star movie rating; day of the week (1: Monday, …, 7: Sunday)

Part 2 solutions are in NTULearn Class-specific Site. As there are many possible correct answers, only a sample of good work presented by students are uploaded.