Exercise 4.1

Part 1. Concepts

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

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

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

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.]

Part 2. Learning Activity: Gender Discrimination Lawsuit

Dataset: Lawsuit.csv

Source: https://www.kaggle.com/hjmjerry/gender-discrimination/home

Female doctors at Houston College of Medicine claimed that the College has engaged in a pattern and practice of discrimination against women in giving promotions and setting salaries.

1. Individually, read the lawsuit documentation PDF and analyze the data using summary table(s) and suitable visualization. Write your own Rscript. You may use data.table, Base R, ggplot2 or other packages.

   A. As an Analytics consultant for **the female doctors**, produce (a) summary tables, and/or (b) charts that show discrimination exists. [More than one correct answer.]

   B. As an Analytics consultant for **the college**, produce (a) summary tables, and/or (b) charts that show discrimination exists. [More than one correct answer.]

Notes:

- There are many possible answers and different ways of presenting the data.
- Instructors will ask for volunteers to present their work during class and will shortlist a few students who had good but different results to present their work.
- Students who present their work in class will gain additional class participation points and/or individual presentation points.
- Presenting students will need to email your work including Rscript to your class instructor to gain the points.
- Class instructor may upload your work (presented in class) into NTULearn class-specific site as a sample for students in the class to refer.