Exercise 4: Data Preprocessing

In the following exercises we are going to use the Dirty Income dataset. You can download this dataset from D2L – Exercise 4.

1. Reading and manually checking.

a. View the file in a text-editor to determine its format and read the file into Python or R.

b. Calculate the number and percentage of observations that are complete.

2. Checking with rules

Besides missing values, the data set contains errors. We have the following background knowledge:

- All employees are adults (18 years old and older).
- All employees pay a tax of 15% of their income.
- All employees make money. The company doesn't hire volunteers who work for free.

What percentage of the data has no errors (i.e., rows that don't violate the above rules)?

3. Correcting

A. Replace non-Female/Male values from **Gender** attribute to either Male or Female.

B. Replace non-positive **income** values to NA.

- optional: can we derive the correct income value from the **tax** field rather than replace it with NA? if you think so, do it!

C. Replace erroneous **Tax** values to NA

- optional: can we derive the correct tax value from **income** field rather than replace it with NA? if you think so, do it!

4. Imputing (inserting missing values)

Use machine learning (e.g., kNN imputation (VIM)) to impute all missing values (replacing NA values with the most predicted values).

Show/take a screenshot of the summary of the data before and after the imputation (check screenshot example)

```
Min.   : 0.000   Min.   :-3.000   Min.   : 0.00
1st Qu.: 5.100   1st Qu.: 2.800   1st Qu.: 1.60
Median : 5.750   Median : 3.000   Median : 4.50
Mean   : 6.559   Mean   : 3.391   Mean   : 4.45
3rd Qu.: 6.400   3rd Qu.: 3.300   3rd Qu.: 5.10
Max.   :73.000   Max.   :30.000   Max.   :63.00
NA's   :10       NA's   :17       NA's   :19
```

Figure 1. Data before kNN imputation

Turn ins:

1. Your calculations and screenshots showing the answers (e.g., percentage of the data that has no errors)
2. The script (R or Python) that was used to load, correct, and impute the dataset.
3. Write a paragraph explaining the importance of cleaning a dataset before providing further analytics about the data.