ASSIGNMENT 2 – Observations, variables, and data matrices

In this assignment we will be performing analysis of the survey to study the smoking habits of UK residents. The answers to the questions asked in the answer is provided below.

(1) What does each row of the data matrix represent?

Ans :

Each row of the data matrix represents a person living in UK. The columns store important information regarding each person such as gender, age, martial status, nationality, ethnicity, gross income, whether they smoke or not etc. In all the data matrix has 1691 rows and 12 columns.

(2) How many participants were included in the survey?

Ans :

There were 1691 participants included in the survey.

(3) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

Ans :

The variables and their corresponding type are given below.

Variable	**Type**
gender	Categorical
age	Numerical – Discrete
maritalStatus	Categorical
highestQualification	Categorical – Ordinal
nationality	Categorical
ethnicity	Categorical
grossIncome	Categorical – Ordinal
region	Categorical
smoke	Categorical
amtWeekends	Numerical – Discrete
amtWeedays	Numerical – Discrete
type	Categorical

(4) According to the CAY classification discussed in the Jacoby reading, how many modes does this data have? How many ways?, How many levels are there for each way?

Ans :

According to CAY classification, this data has two modes, the modes are survey respondents and survey items. It has two ways one each for the respondents and items and respondents have 1691 levels and items have 12 levels.

(5) Consider this proposal: if the survey had asked for "years of formal education" instead of "highest qualification," then we could interpret the responses as values of a ratio level variable. Take a position

on whether that interpretation would be reasonable and useful. Explain and justify your argument in no more than two paragraphs.

Ans :

Although "years of formal education" would have the comparative qualities of "highest qualification" and also the benefits of numeric variables such as we would be able to calculate the mean and median of the values, it would also add more complexity into the dataset by not capturing some background information. For example, if a person who has a bachelor's degree had to repeat few years to get bachelor's degree due to backlogs and other issues would have his years of formal education increased by come amount. If that number is two or more he would be on the same category as a masters student who didn't have to repeat his years if this new variable is used. This background information is not captured here. Also, the number of years to get a PhD also differ from person to person and even though they have the same degree level the years variable would have different values for them. All these examples show that a greater number of years does not convey more education. So, a ratio level interpretation would not be reasonable and useful.

(6) Propose two different interpretations of the gross income variable, following the four-way Coombs classification discussed in the Jacoby reading. In which two categories would observations on income be classified. Explain your reasoning.

Ans :

The categories in which observations on gross income variable can be are Single Stimulus and Stimulus Comparison. The single stimulus category can be applied as the respondent can be said to have a dominance relationship with the income value. The respondent can be less than, between or greater than the income value. So, the respondent dominates all the units of income below its range. For e.g. if the respondent has an income between 10,400 to 15,600, it can be said that it dominates all the units below 10,400. On the other hand, the observations on the income can be used to compare and order the income levels among themselves. So, an income range 10,400 to 15,600 is below 15,600 to 20,800 but above 5,200 to 10,400. The geometric model for both the interpretations is below.

Single Stimulus – Respondent X has income between 10,400 and 15,600

```
              X
    -------------------------
10,400              15,600
```

Stimulus Comparison – Income range 10,400 to 15,600 (A) is below 15,600 to 20,800 (B) and above 5,200 to 10,400 (C)

```
         C      A      B
    ---------------------------
```