Exercise Unit 4

In the Unit 4 lecturecast, students were asked to perform an exercise presented in the "Data Wrangling with Python" book. Although the exercise specified the pages, I could not find them because the book was available online. I searched for the terms "mn.csv" and "mn_header.csv," and the search led me to Chapter 7. I then followed that example to complete the exercise.

For the exercise, I downloaded the "Zimbabwe MICS6 SPSS Datasets" and chose to use the document called "mn.sav". Initially, I uploaded the SPSS dataset and asked the program to show me the first few rows and technical column names. I decided to reduce the dataset to the first ten columns and rename only those.

To rename them, I created a CSV file with two columns: code (containing the original column names) and description (containing the full name of each column). The description of each variable was obtained from the World Bank website (https://microdata.worldbank.org/catalog/4180/data-dictionary/F7?file_name=mn.sav)

I uploaded the file I created, called "mn_headers.csv," to Python and replaced the technical names with the full variable names.

Finally, I saved a new CSV file with the reduced dataset and read it into a list of dictionaries, where each dictionary represented a row, with column names as keys and the corresponding data as values.

To continue the exercise, I started the cleaning phase, as proposed in Chapter 7. First, I removed all duplicates and created a new list with unique rows. This step was only

performed to follow the exercise faithfully. In reality, there were no duplicate entries in this dataset.

I then grouped the data by family by combining the cluster number and household number. I created a dictionary in which each key represents a family, and the associated value is a list of family members. The result was 3,588 keys, indicating that there were 3,588 unique families and that some families had more than one person in the dataset.

Finally, I calculated the average number of family members per household, which was 1.3. This means that, in the majority of the households, there was only one member, while in others, there was more than one member.