**Homework 1 Report**

**Data Analysis and Machine Learning with Python**

**B09901061 莊政達**

**Dataset 1: Housing**

**A1.**

1. Check if there is duplicated data and delete the duplicates.
2. Check if there is a missing value in the data. Handle the missing value according to a determined strategy. For example, delete the whole row or column if there is a missing value, replace the missing value with other data’s mean, or simply fill the box with a “NAN” word.

**A2.** I received the following response from ChatGPT:

*“ 1. Handle Missing Values:*

*- Identify missing values in the dataset.*

*- Decide on a strategy to handle missing values, such as imputation (replacing missing values with a calculated or estimated value), deletion of rows or columns with missing values, or considering missing values as a separate category.*

*2. Check for Duplicates:*

*- Identify and remove duplicate rows from the dataset to avoid bias in the analysis.*

*3. Data Formatting:*

*- Standardize formats for consistency, such as date formats, numeric formats, and text cases.*

*4. Data Transformation:*

*- If necessary, transform variables by applying transformations like log transformation, scaling, or encoding categorical variables using techniques such as one-hot encoding or label encoding.*

*5. Normalize or Standardize Data:*

*- Scale numerical features if needed to ensure they have similar ranges, especially when using algorithms sensitive to feature scales like gradient descent-based algorithms.*

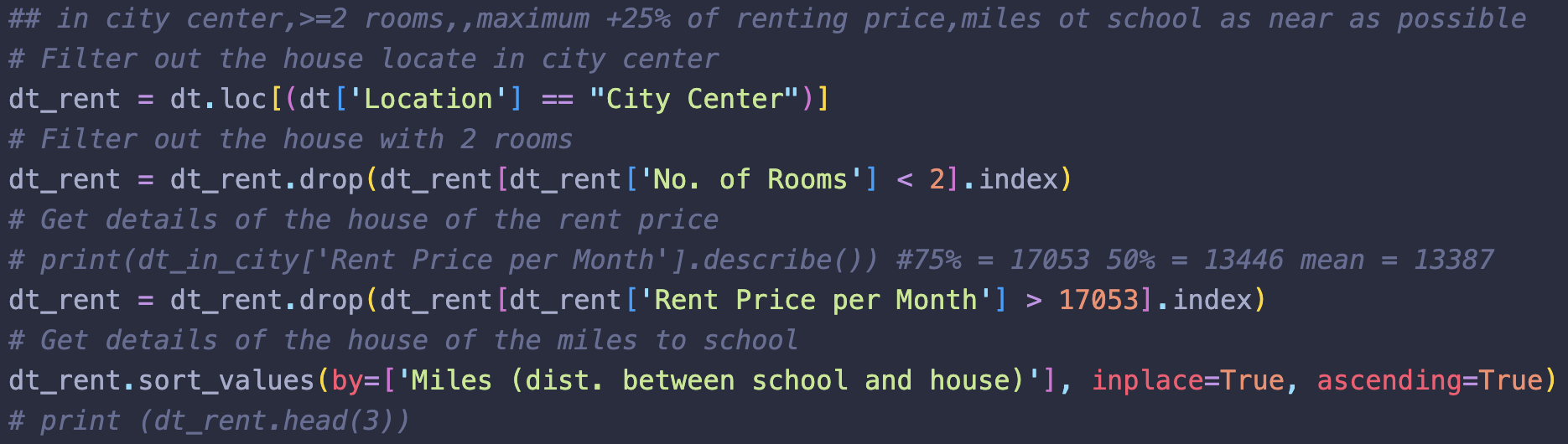
*These steps are fundamental in ensuring that the data is clean, consistent, and ready for further analysis without introducing biases or errors due to missing values, duplicates, or inconsistent formats.”*

*Ref: https://chat.openai.com/share/217c7ea7-8625-4f32-84fb-419b8f40c85d*

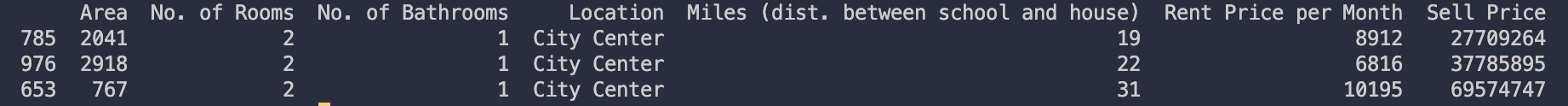
The answers are reasonable since ChatGPT not only mentions handling missing values and duplicates but also more pre-processing to make the data more readable such as formatting.

However, it’s worth mentioning that data transformation isn’t suitable for every data. When data’s uniqueness is the specialist, we should rather abandon the whole row or column than modify the value.

**A3.** According to the given data, I will set some conditions to decide which ones to rent. The following are the conditions that must be met:

1. Located in the city center.
2. More than 1 room.
3. The renting price must not exceed 75% of the available choices.
4. As near as possible to schools.

*Figure 1: Filter out each condition*

After analyzing the data, I prefer to choose 1 of the following 3 houses to rent, which is index 785,976, or 653 (in preference order).

*Figure 2: Dataset 1 A3 result*

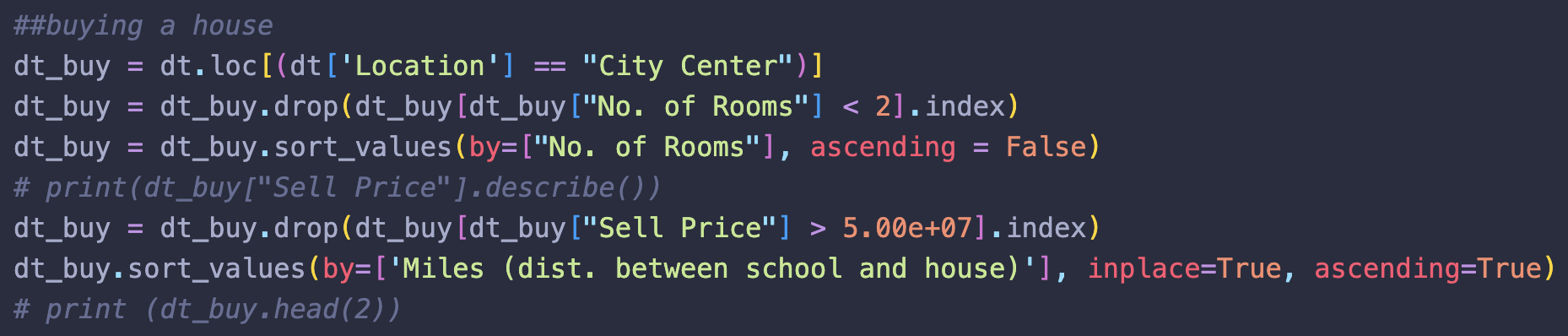
**A4.** If I have enough funds to purchase a house, I would proceed with a purchase for sure. Just like choosing a house to rent, my conditions that must be met are as follows:

1. Located in the city center.

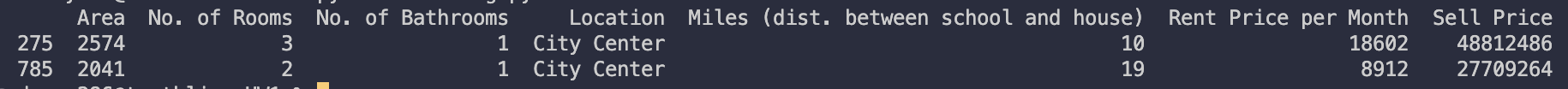
2. More than 1 room.

3. The renting price must not exceed 5.00e+07 (between 50% to 75% of the available choices)

4. As near as possible to schools.

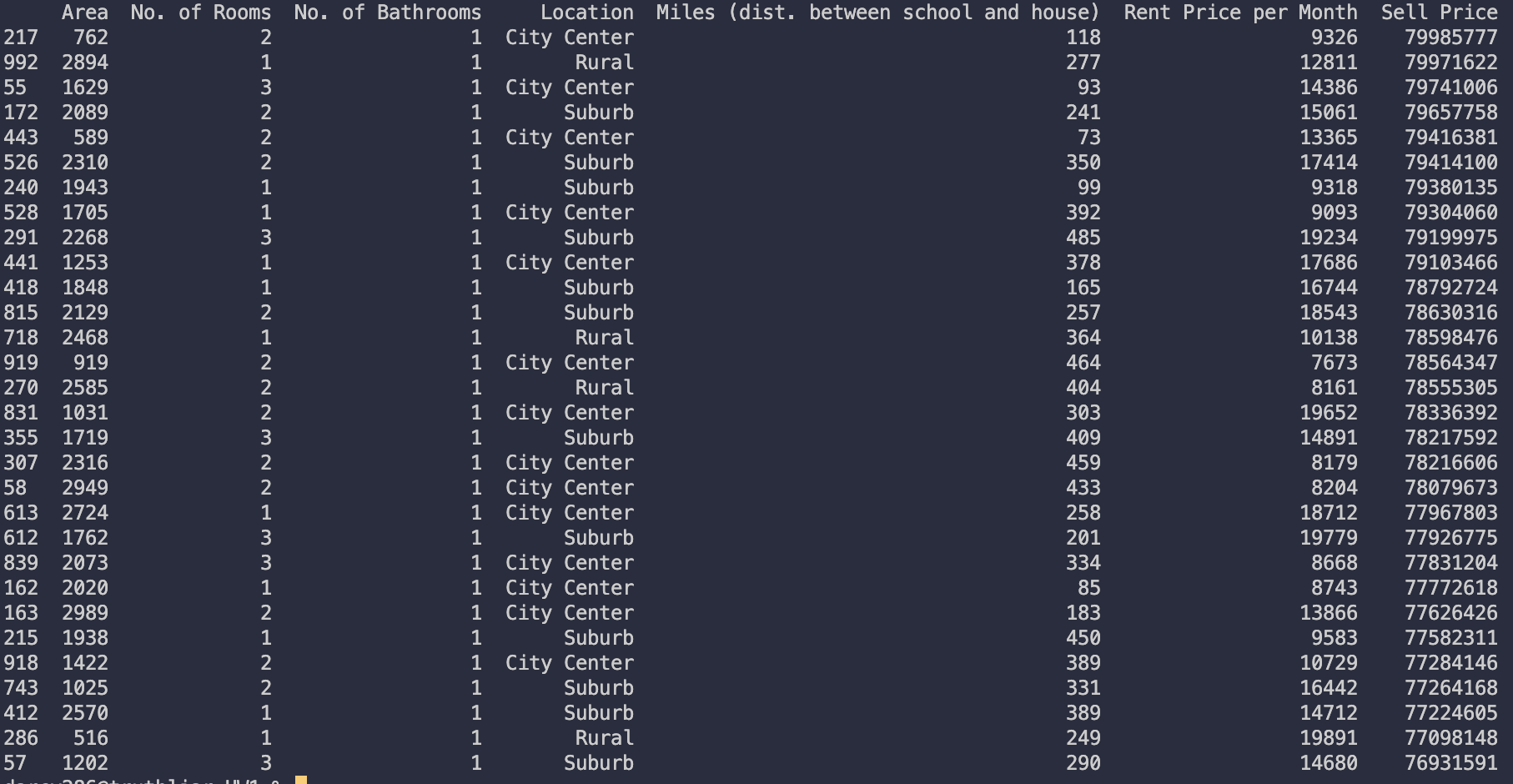


*Figure 3: Filter out each condition*

After analyzing the data, I prefer to choose 1 of the following 2 houses to buy according to my budget. If I’m not tight on budget, I will purchase the house that is in index 275, otherwise, I will buy the house that is in index 785.

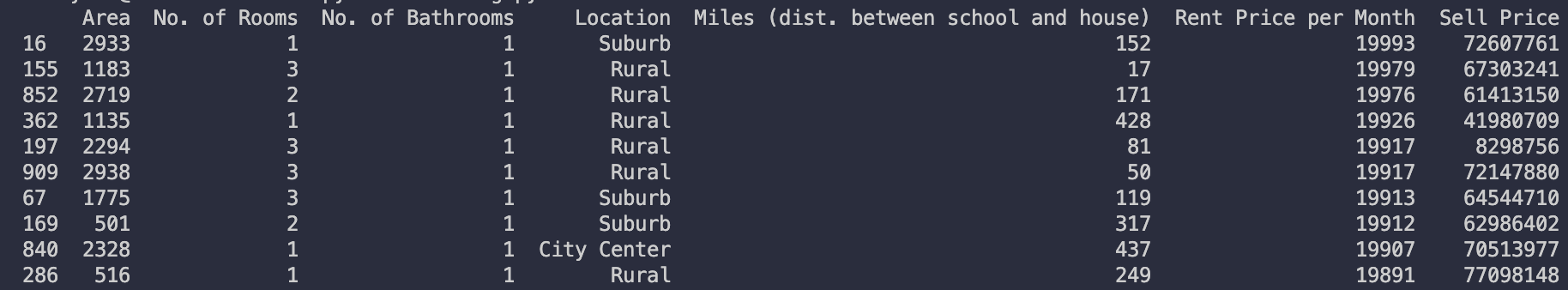
*Figure 4: Dataset 1 A4 result*

**A5.**

**For Sale:** House 286 is unusually high. It is located in rural with 1 room and 1 bathroom. The area is only 516, which is much lower than the average area of houses in the data. However, it cost 77098148 and exceeds 75% of houses in the data.

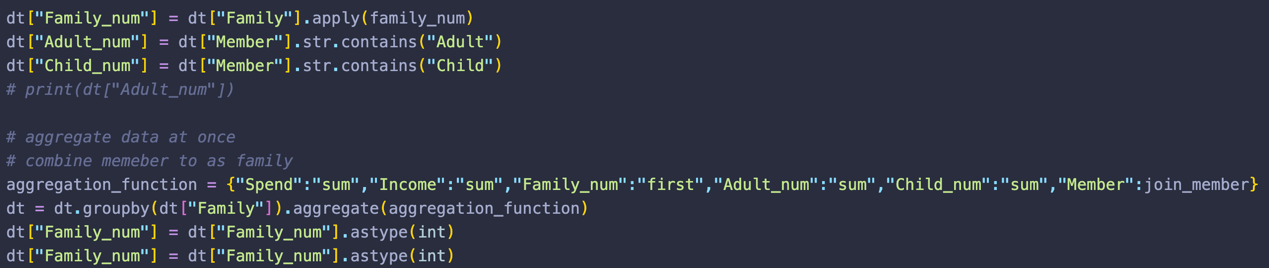
*Figure 5: Top 20 highest-selling prices in data*

**For Rent:** House 169 is unusually high. It is located in the suburb with 2 rooms and 1 bathroom. The distance to a school is 317 miles, which isn’t a short distance. However, the rent per month is 19912, which is the eighth highest rental price in the data. The price exceeds many houses that have better locations or bigger areas.



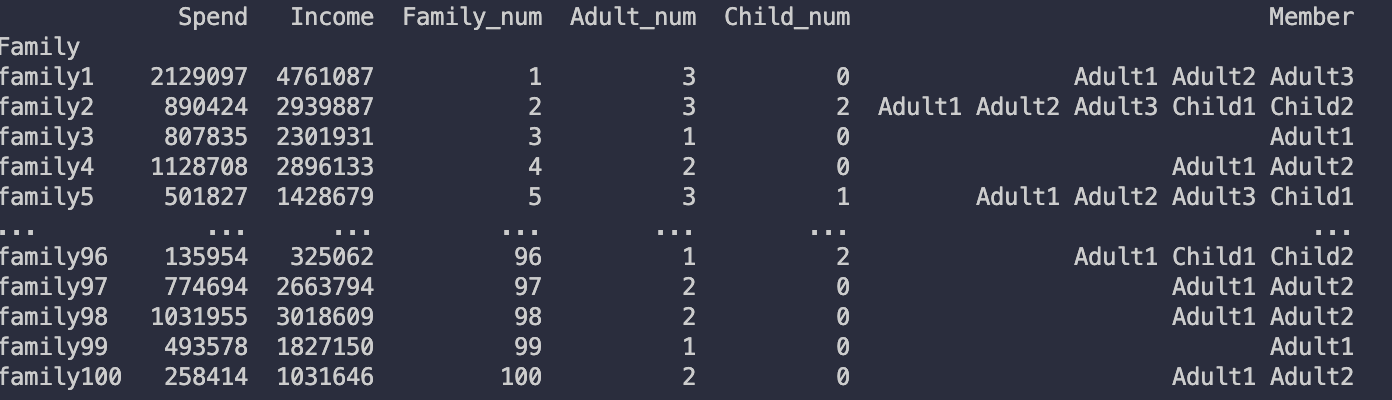
*Figure 6: Top 10 highest rent prices in data*

**Dataset 2: Family**

**A1.** I group the data by family to get the value of each family’s total income and spending. Starting from adding columns to count family, adults, and children.

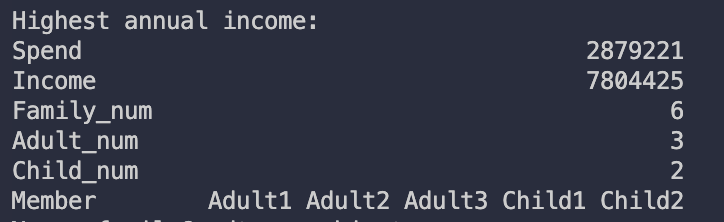
*Figure 7: Process to aggregate the data*

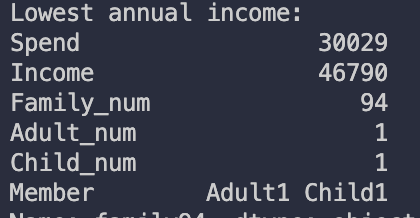
And then aggregate the data to create a more user-friendly data frame.

The example result is as follows

*Figure 8: Dataset 2 A1 result*

Sort the data by total income, we know Family 6 has the highest annual income, which is 7804425. And Family 94 has the lowest annual income, which is 46790.



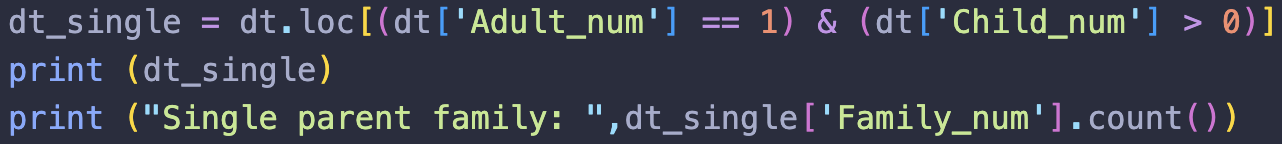
*Figure 9: Family that has the highest annual income*

*Figure 10: Family that has the lowest annual income*

**A2.** According to the given data, no family doesn’t possess adequate annual income to cover all members’ spending. I determine this by searching if there is a family’s spending > income and no result is found.



*Figure 11: Dataset 2 A2 result*

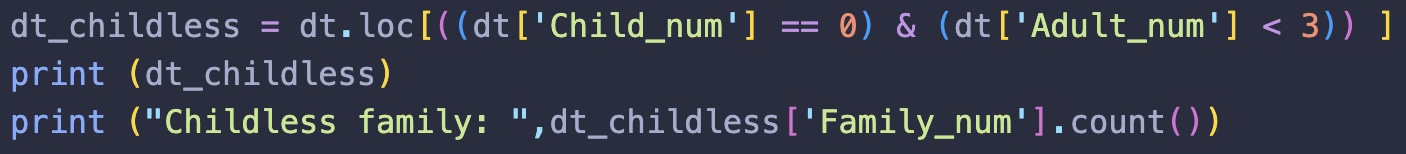
**A3.** Yes, there are 23 single-parent families and 27 childless families. To count single-parent families, I calculate the family with at least 1 child but only 1 parent at present.

*Figure 12: Functions filtering out single-parent family*

**

*Figure 13: Dataset 2 A3 result – Single-parent family*

For childless families, I take the families with no children and the number of adults is less than 3.



*Figure 14: Functions to filter out single-parent family*

**

*Figure 15: Dataset 2 A3 result – Childless family*

**A4.** I don’t see obvious errors like missing or duplicate data. However, I think the data could also include the age of the members, which could be more accurate when calculating information like single-parent families. For example, if a family has only 2 adults can’t be determined whether it is a parent and grown children or a couple.

**A5.** I think ChatGPT or Bing can assist with the four questions above. With AI tools given suitable prompts and files, I believe it can solve the four questions completely. Besides the questions above and the files, I will also prompt it with more specific conditions like a single-parent family means there is only 1 parent taking care of at least 1 child, and a childless family doesn’t include if there is a child who grows up and becomes an adult currently.