Report of HW1
B09602017 白宗民

Dataset 1: Housing Data
Q1. What steps will you take upon receiving this dataset before commencing data analysis?

1. **Search** the indicators about the housing decision.

2. **Cleaning** the data by handling the useless data or data missing.

3. **Deleting** the column [‘No. Of bathroom’] since I found that it’s all the same.

4. Change the numeric part from **string** to **float** the further calculation.

5. **Adding** the column searched on the internet via python implementation.

6. **One-hot encoding** the “Location” before ML training process.

- It’s all about adding and removing the columns or checking the data completeness.

Q2. If you are to inquire about Q1 from ChatGPT or Bing, what responses will you receive? Do you find them reasonable? If not, how will you rectify it?

1. Data Cleaning:

   - Check for any missing or null values in the dataset and decide how to handle them (e.g., imputation or removal).

   - Look for any inconsistencies or errors in the data (e.g., outliers, incorrect entries) and correct them if possible.

2. Data Understanding:

   - Understand the meaning of each column in the dataset (e.g., "Area", "No. of Rooms", "Rent Price per Month”).

   - Clarify any ambiguous or unclear data points with the provider of the dataset.

3. Data Exploration:

   - Conduct exploratory data analysis (EDA) to understand the distribution of variables, identify patterns, and detect outliers.

   - Visualize the data using plots such as histograms, box plots, scatter plots, etc., to gain insights into the relationships between variables.

4. Feature Engineering:

   - Create any additional features that might be useful for analysis (e.g., calculating the price per square foot, categorizing locations into different regions).

- Encode categorical variables if necessary (e.g., converting "Location" from text to numerical values using one-hot encoding).

5. Normalization/Standardization:

   - Depending on the analysis method to be used, normalize or standardize numerical features if needed to bring them to a similar scale.

6. Data Splitting:

   - If the dataset is intended for predictive modeling, split it into training and testing sets to evaluate the performance of the model accurately.

7. Define Analysis Goals:

   - Clearly define the objectives of the data analysis and the questions to be answered using the dataset.

By following these steps, I can ensure that the dataset is properly prepared and understood before beginning the analysis, leading to more reliable and meaningful insights. - Generated by **chatGPT-3.5**

Q3. If you are restricted to renting a house, which one or ones will you select, and why?

Actually I think this depends on the financial condition and the needs at the moment.

Assume that I am a person who likes to live at a convenient area, which is **city center**.

And the **average rent price** at city center is **13272.215909**

1. If I am single and have few salary: **House 326;** since it has the 2nd lowest rent price per month, big area with only one room(single situation) and only 50 miles distance.

2. If I am not single and have few salary: **House 976**; since it has the 9th lowest rent price and has 2 rooms with 2918 area for lovely family or couple and it’s also 22 miles distance.

3. If I am single and have a lot of salary: **House 273;** since it has the largest area with only 55 miles distance and only one room for single Vincent.

4. If I am not single and have a lot of salary: **House 976**; since it has the 7th largest area and has 2 rooms for lovely family or couple and it’s also 22 miles distance. Not to mention that it has the 9th lowest rent price per month.

All the above analysis could be seen in my script’s part of Q3.

Q4. Assuming you have enough funds to purchase a house, will you opt to continue renting or proceed with a purchase? If renting, which one will you choose? If buying, which one will you select? Why?

Definitely **buy** the house. And I would choose the one with (# of room > 1), (Area>2000) and (Distance < 200) since buying a house is a investment for future family.

So I would choose **House 768** since it’s the one that satisfied all the condition and has the lowest sell price. And it’s ROI is 9.03% so maybe I can rent it someday in the future if I earn more money and can buy a new fancy house.

Q5. Are there any properties with rent or selling prices that seem unusually high or low? Why?

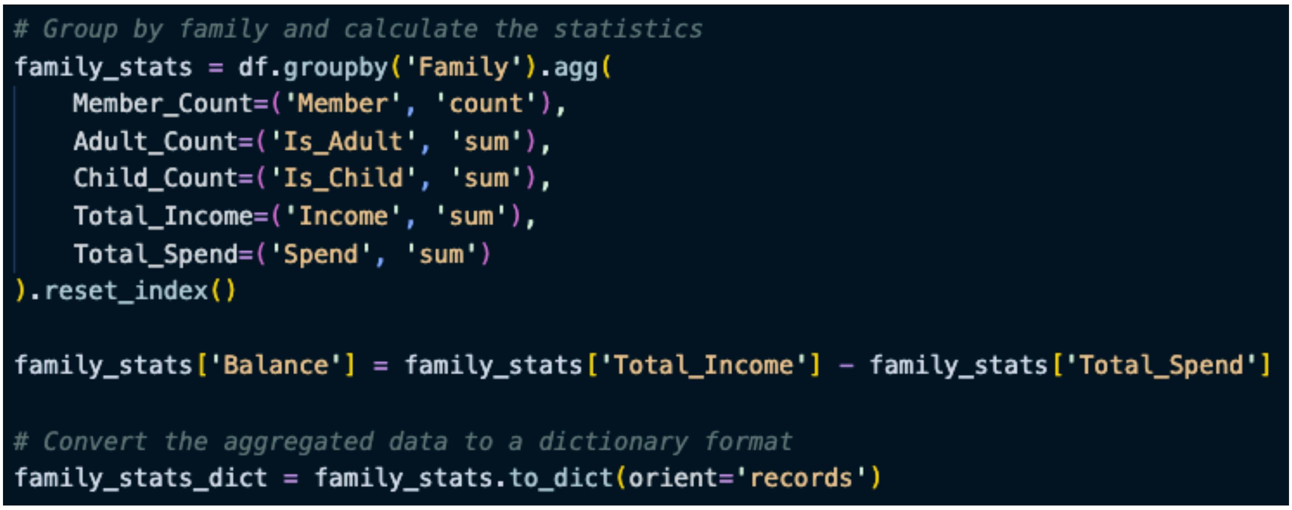
For Rent: I think that **House 362** is unusually high: since it has only 1 room with 1135 area and it’s located in rural with also 428 miles distance. But it’s price is the 4th highest for the rent price per month

For Sell: I think that **House 286** is unusually high: since it has only 1 room with 516 area and it’s located in rural with also 249 distance. But it cost 77098148 which is the 4th highest set price for rural house and it’s only 2887629 lower than the house with highest sell price.

(All the house index is the same with the index in the css file, which is started from 0)

Dataset 2: Family Data

At first, I construct a **list** where each element is a **dictionary** with each family, as a result, this list has 100 elements.
Each dictionary has keys: **Family ID, Member_count, Adult_count, Child_count, Total_income, Total_spend, Balance.**

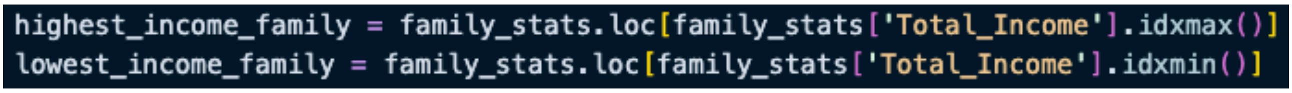


```
# Group by family and calculate the statistics
family_stats = df.groupby('Family').agg(
    Member_Count=('Member', 'count'),
    Adult_Count=('Is_Adult', 'sum'),
    Child_Count=('Is_Child', 'sum'),
    Total_Income=('Income', 'sum'),
    Total_Spend=('Spend', 'sum')
).reset_index()

family_stats['Balance'] = family_stats['Total_Income'] - family_stats['Total_Spend']

# Convert the aggregated data to a dictionary format
family_stats_dict = family_stats.to_dict(orient='records')
```

Q1. Which family boasts the highest annual income, and which has the lowest? How do you ascertain this?



```
highest_income_family = family_stats.loc[family_stats['Total_Income'].idxmax()]
lowest_income_family = family_stats.loc[family_stats['Total_Income'].idxmin()]
```

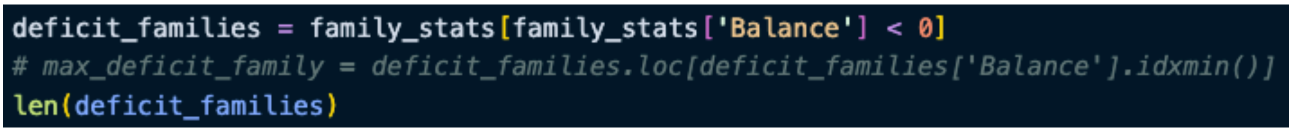
Highest income: **family 6** with 7804425

Lowest income: **family 94** with 46790

By the code I can ascertain it.

Q2. Which families do not possess adequate annual income to cover all members' spending? What is the maximum shortfall? How do you determine this?

There is **no** family does not possess adequate annual income to cover all members' spending. Determine by the code:

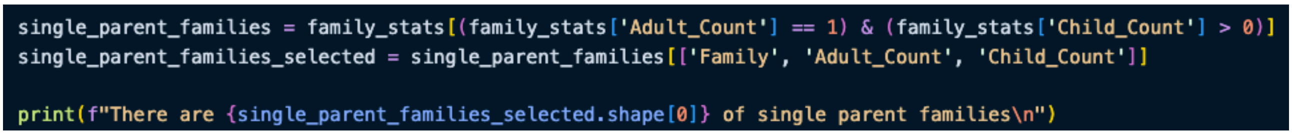


```
deficit_families = family_stats[family_stats['Balance'] < 0]
# max_deficit_family = deficit_families.loc[deficit_families['Balance'].idxmin()]
len(deficit_families)
```

With **len(deficit_family) = 0**

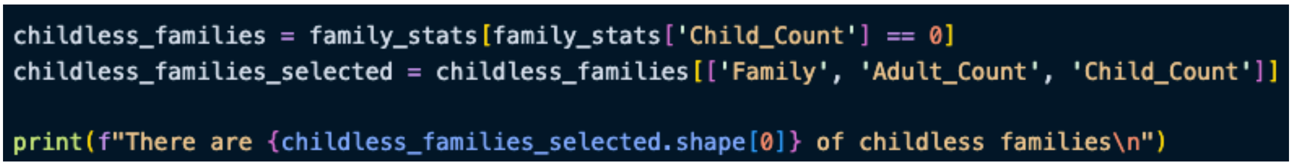
Q3. Are there any single-parent families, where only one Adult is present? Are there any childless families? How do you discern this?

Since I construct the **bool** column: **is_adult** and **is_child**, by the code below I can discern.



```
single_parent_families = family_stats[(family_stats['Adult_Count'] == 1) & (family_stats['Child_Count'] > 0)]
single_parent_families_selected = single_parent_families[['Family', 'Adult_Count', 'Child_Count']]

print(f"There are {single_parent_families_selected.shape[0]} of single parent families\n")
```



```
childless_families = family_stats[family_stats['Child_Count'] == 0]
childless_families_selected = childless_families[['Family', 'Adult_Count', 'Child_Count']]

print(f"There are {childless_families_selected.shape[0]} of childless families\n")
```

Where the **single-parent** families: 23

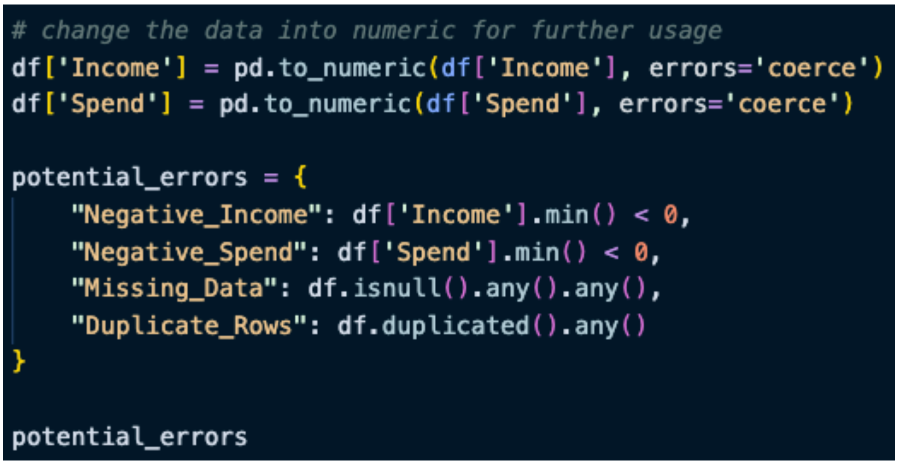
And the **childless** families: 35

Q4. Do you suspect any errors within this dataset? Examples may include negative figures, missing or duplicate data, etc. Why?

Actually nah by the code:

I checked the numeric number should be positive and check the string should not be null and the duplicated rows also.

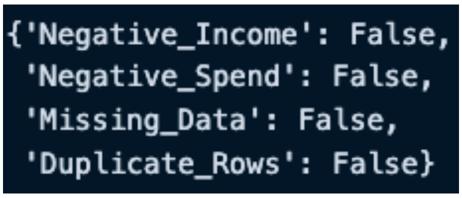
So the data would not be **negative** (not null for sure),not **null** and not **duplicated**.



```
# change the data into numeric for further usage
df['Income'] = pd.to_numeric(df['Income'], errors='coerce')
df['Spend'] = pd.to_numeric(df['Spend'], errors='coerce')

potential_errors = {
    "Negative_Income": df['Income'].min() < 0,
    "Negative_Spend": df['Spend'].min() < 0,
    "Missing_Data": df.isnull().any().any(),
    "Duplicate_Rows": df.duplicated().any()
}

potential_errors
```



```
{'Negative_Income': False,
 'Negative_Spend': False,
 'Missing_Data': False,
 'Duplicate_Rows': False}
```

Q5. Can ChatGPT or Bing assist with the aforementioned four questions? If so, to what extent? How do you issue commands to the AI tool? If not, why not?

This is generated by chatGPT-3.5

Yes, but I like the code I wrought more.

To answer your questions, I'll first analyze the provided dataset:

- **Highest and Lowest Annual Income:**
    - To find the family with the highest and lowest annual income, I'll sum up the incomes for each family and then compare them.

- **Families with Inadequate Annual Income:**
    - I'll calculate the total spend for each family and compare it with their total income to identify families with inadequate income. The maximum shortfall would be the largest deficit among these families.
- **Single-parent Families and Childless Families:**
    - I'll identify single-parent families by checking for families with only one member listed as an adult. Childless families would be those without any child members.
- **Detecting Errors in the Dataset:**
    - I'll look for negative incomes or spending, missing data, or duplicate entries.

My prompt:

Q1. Which family boasts the highest annual income, and which has the lowest? How do you ascertain this?
Q2. Which families do not possess adequate annual income to cover all members' spending? What is the maximum shortfall? How do you determine this?
Q3. Are there any single-parent families, where only one Adult is present? Are there any childless families? How do you discern this?
Q4. Do you suspect any errors within this dataset? Examples may include negative figures, missing or duplicate data, etc. Why?

data format:
Family,Member,Income,Spend
family1,Adult1,2376330,1119433
family1,Adult2,130268,37337
family1,Adult3,2254489,972327
family2,Adult1,2292355,649806
family2,Adult2,298167,100723
family2,Adult3,349365,134693
family2,Child1,0,1769
family2,Child2,0,3433
family3,Adult1,2301931,807835
…

This is very important to my career (from *LLM Stimuli*), please think step by step (from *CoT*).

Reference

https://arxiv.org/abs/2307.11760
https://arxiv.org/abs/2201.11903
https://chat.openai.com/c/d5d78661-725d-4be6-9306-991d86d7d944
https://rich01.com/rental-yield-calculation/