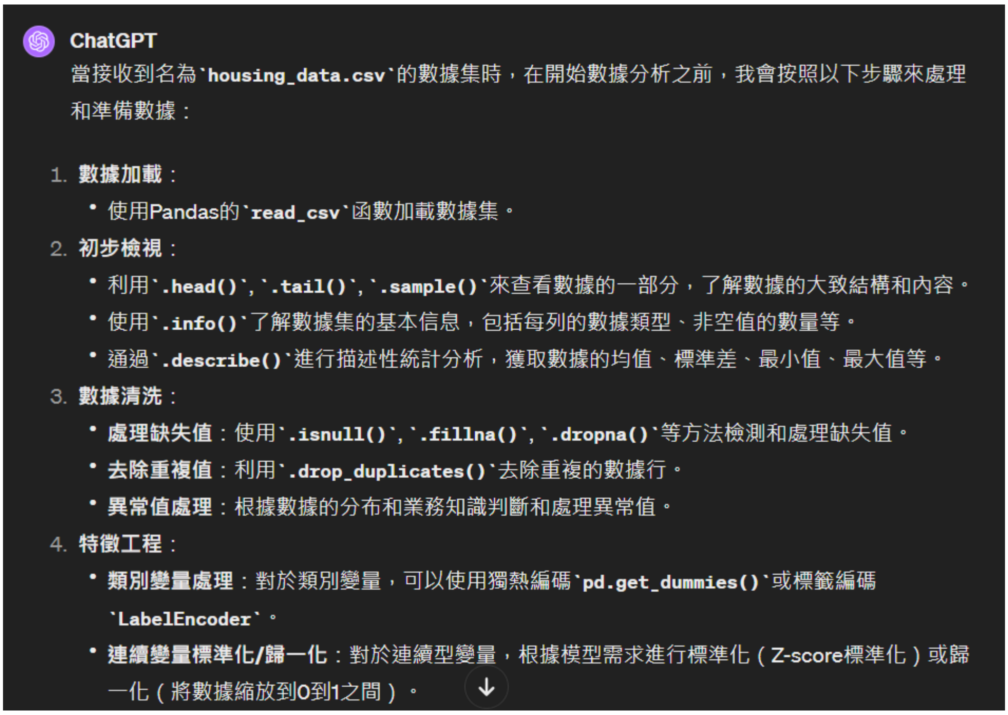
Dataset 1: housing_data.csv

Q1. What steps will you take upon receiving this dataset before commencing data analysis?

檢查資料形狀及欄位名稱，確認內容是否與問題有關。
有無NaN值。
有無重複值。
確認NaN值跟重複值是否需要刪除或是填充。

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?

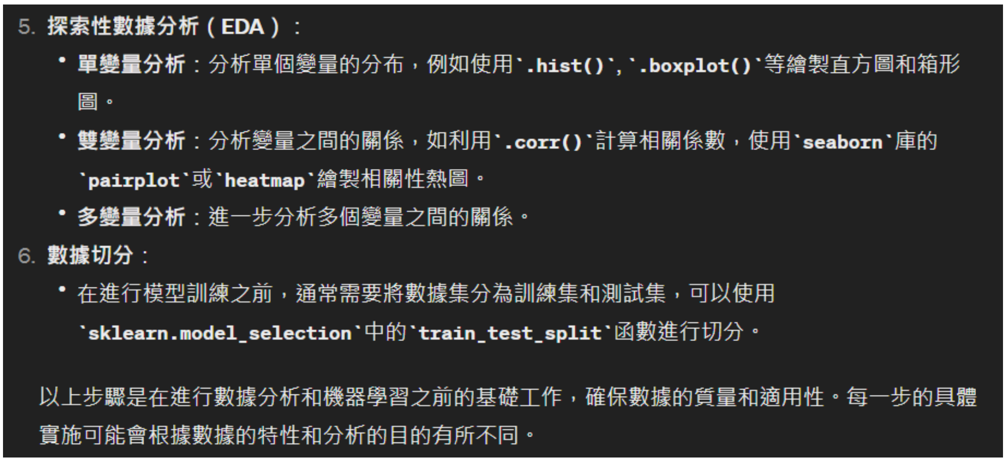
我覺得答覆合理，ChatGPT檢查的比我還仔細，同時又讓我知道更多處理方法與細節。



ChatGPT

當接收到名為`housing_data.csv`的數據集時，在開始數據分析之前，我會按照以下步驟來處理和準備數據：

1. **數據加載**：
   - 使用Pandas的`read_csv`函數加載數據集。
2. **初步檢視**：
   - 利用`.head()`, `.tail()`, `.sample()`來查看數據的一部分，了解數據的大致結構和內容。
   - 使用`.info()`了解數據集的基本信息，包括每列的數據類型、非空值的數量等。
   - 通過`.describe()`進行描述性統計分析，獲取數據的均值、標準差、最小值、最大值等。
3. **數據清洗**：
   - **處理缺失值**：使用`.isnull()`, `.fillna()`, `.dropna()`等方法檢測和處理缺失值。
   - **去除重複值**：利用`.drop_duplicates()`去除重複的數據行。
   - **異常值處理**：根據數據的分布和業務知識判斷和處理異常值。
4. **特徵工程**：
   - **類別變量處理**：對於類別變量，可以使用獨熱編碼`pd.get_dummies()`或標籤編碼`LabelEncoder`。
   - **連續變量標準化/歸一化**：對於連續型變量，根據模型需求進行標準化（Z-score標準化）或歸一化（將數據縮放到0到1之間）。

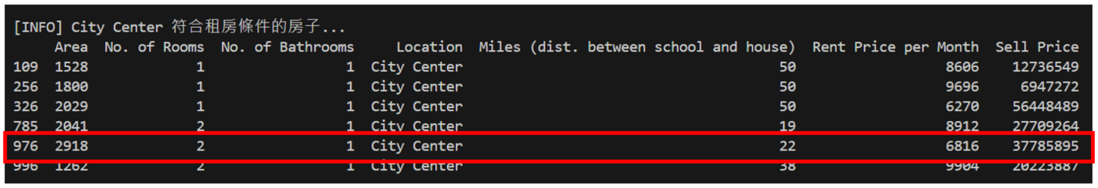


5. **探索性數據分析（EDA）**：
   - **單變量分析**：分析單個變量的分布，例如使用`.hist()`, `.boxplot()`等繪製直方圖和箱形圖。
   - **雙變量分析**：分析變量之間的關係，如利用`.corr()`計算相關係數，使用`seaborn`庫的`pairplot`或`heatmap`繪製相關性熱圖。
   - **多變量分析**：進一步分析多個變量之間的關係。
6. **數據切分**：
   - 在進行模型訓練之前，通常需要將數據集分為訓練集和測試集，可以使用`sklearn.model_selection`中的`train_test_split`函數進行切分。

以上步驟是在進行數據分析和機器學習之前的基礎工作，確保數據的質量和適用性。每一步的具體實施可能會根據數據的特性和分析的目的有所不同。

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

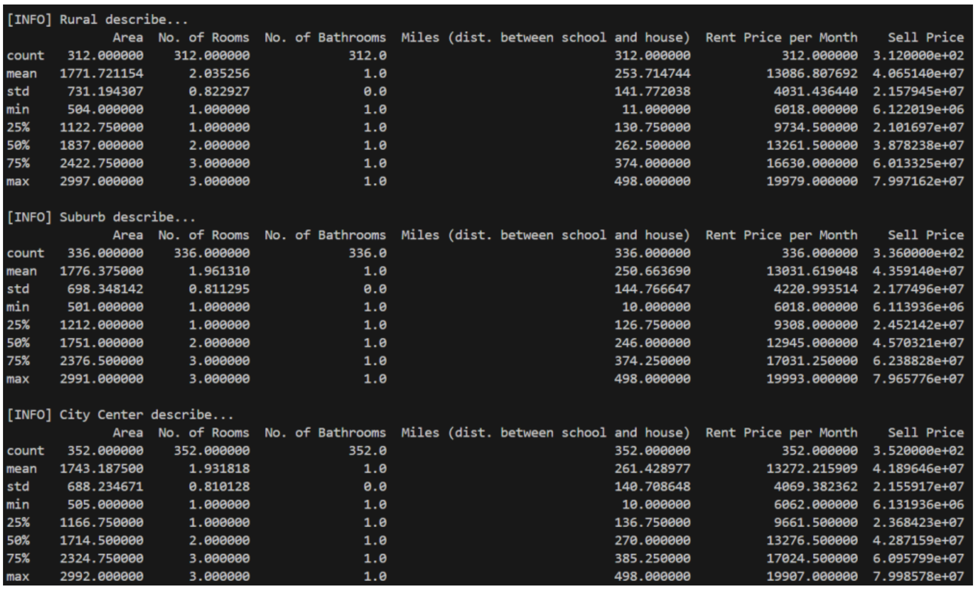
我會選擇租City Center 每月租金6816 距離學校22 miles 面積2918的房子。



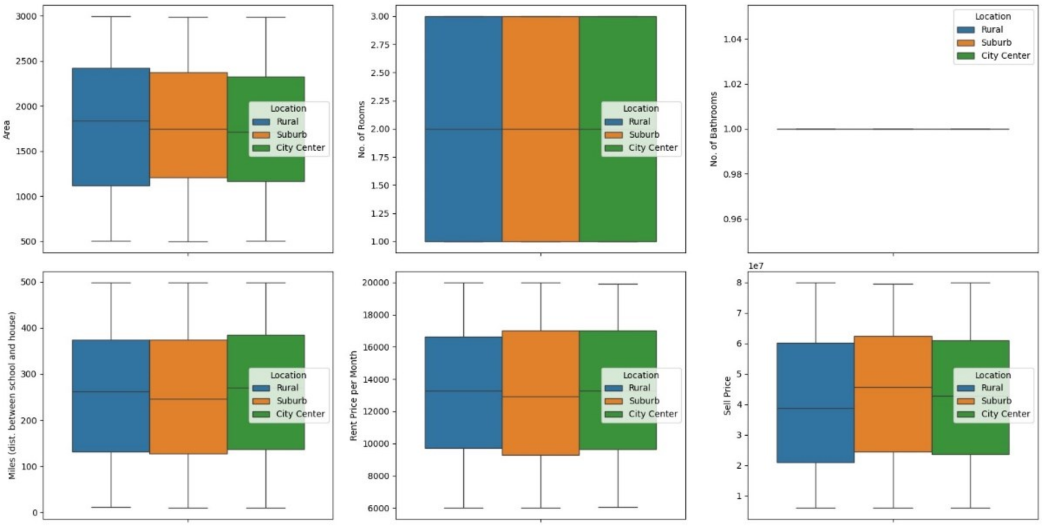
```
[INFO] City Center 符合租房條件的房子...
     Area  No. of Rooms  No. of Bathrooms     Location  Miles (dist. between school and house)  Rent Price per Month  Sell Price
109  1528             1                 1  City Center                                      50                  8606    12736549
256  1800             1                 1  City Center                                      50                  9696     6947272
326  2029             1                 1  City Center                                      50                  6270    56448489
785  2041             2                 1  City Center                                      19                  8912    27709264
976  2918             2                 1  City Center                                      22                  6816    37785895
996  1262             2                 1  City Center                                      38                  9904    20223887
```

因為我想找距離學校近且房租低的房子，以下為篩選的過程及圖表。

1. 以Location將資料拆成三份，打印出它們的describe()：
   在這裡我發現每間房子最少有一間房間跟浴室，所以考量的欄位就剩下距離、房租、面積。

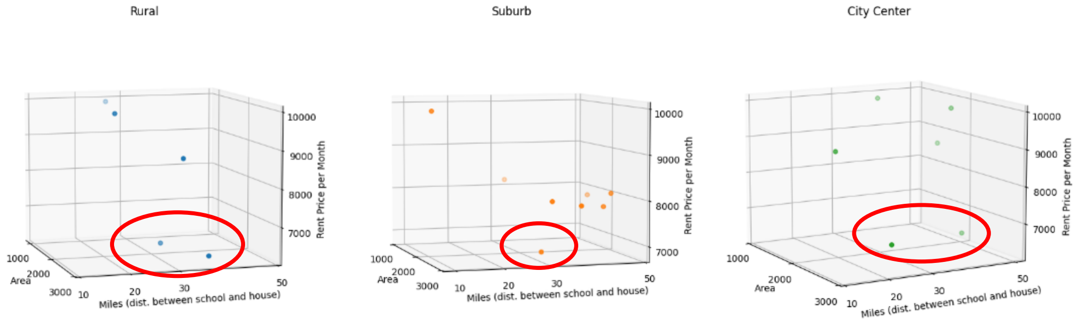


```
[INFO] Rural describe...
              Area  No. of Rooms  No. of Bathrooms  Miles (dist. between school and house)  Rent Price per Month    Sell Price
count   312.000000    312.000000             312.0                              312.000000            312.000000  3.120000e+02
mean   1771.721154      2.035256               1.0                              253.714744          13086.807692  4.065140e+07
std     731.194307      0.822927               0.0                              141.772038           4031.436440  2.157945e+07
min     504.000000      1.000000               1.0                               11.000000           6018.000000  6.122019e+06
25%    1122.750000      1.000000               1.0                              130.750000           9734.500000  2.101697e+07
50%    1837.000000      2.000000               1.0                              262.500000          13261.500000  3.878238e+07
75%    2422.750000      3.000000               1.0                              374.000000          16630.000000  6.013325e+07
max    2997.000000      3.000000               1.0                              498.000000          19979.000000  7.997162e+07

[INFO] Suburb describe...
              Area  No. of Rooms  No. of Bathrooms  Miles (dist. between school and house)  Rent Price per Month    Sell Price
count   336.000000    336.000000             336.0                              336.000000            336.000000  3.360000e+02
mean   1776.375000      1.961310               1.0                              250.663690          13031.619048  4.359140e+07
std     698.348142      0.811295               0.0                              144.766647           4220.993514  2.177496e+07
min     501.000000      1.000000               1.0                               10.000000           6018.000000  6.113936e+06
25%    1212.000000      1.000000               1.0                              126.750000           9308.000000  2.452142e+07
50%    1751.000000      2.000000               1.0                              246.000000          12945.000000  4.570321e+07
75%    2376.500000      3.000000               1.0                              374.250000          17031.250000  6.238828e+07
max    2991.000000      3.000000               1.0                              498.000000          19993.000000  7.965776e+07

[INFO] City Center describe...
              Area  No. of Rooms  No. of Bathrooms  Miles (dist. between school and house)  Rent Price per Month    Sell Price
count   352.000000    352.000000             352.0                              352.000000            352.000000  3.520000e+02
mean   1743.187500      1.931818               1.0                              261.428977          13272.215909  4.189646e+07
std     688.234671      0.810128               0.0                              140.708648           4069.382362  2.155917e+07
min     505.000000      1.000000               1.0                               10.000000           6062.000000  6.131936e+06
25%    1166.750000      1.000000               1.0                              136.750000           9661.500000  2.368423e+07
50%    1714.500000      2.000000               1.0                              270.000000          13276.500000  4.287159e+07
75%    2324.750000      3.000000               1.0                              385.250000          17024.500000  6.095799e+07
max    2992.000000      3.000000               1.0                              498.000000          19907.000000  7.998578e+07
```

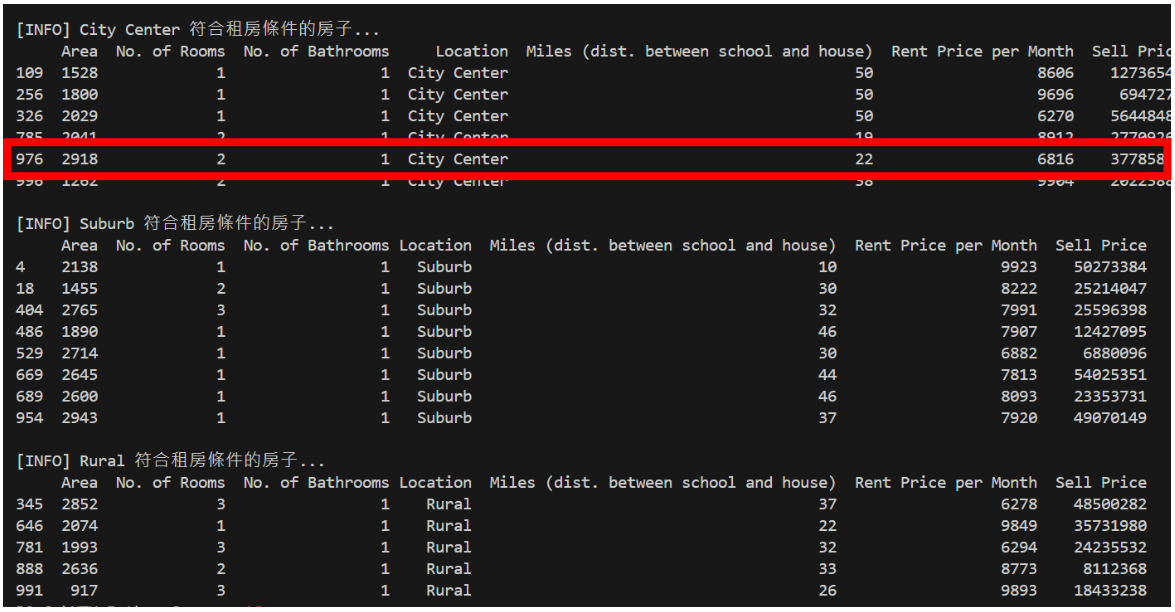


2. 先濾掉資料集Miles大於50的房子，然後濾掉房租大於10000的房子：
看到了幾間價格低、離學校近、面積大的房子。

X軸：房屋面積 Y軸：距離學校 Z軸：每月房租



3. 打印出三個地區符合條件的房子：
最終選擇City Center每月租金6816 距離學校22 miles 面積2918 兩房的房子

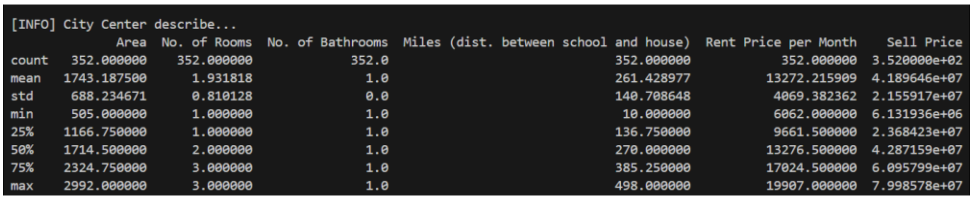


```
[INFO] City Center 符合租房條件的房子...
     Area  No. of Rooms  No. of Bathrooms     Location  Miles (dist. between school and house)  Rent Price per Month  Sell Pric
109  1528             1                 1  City Center                                      50                  8606   1273654
256  1800             1                 1  City Center                                      50                  9696    694727
326  2029             1                 1  City Center                                      50                  6270   5644848
785  2041             2                 1  City Center                                      19                  8912   2770926
976  2918             2                 1  City Center                                      22                  6816    377858
998  1262             2                 1  City Center                                      38                  9904   2022388

[INFO] Suburb 符合租房條件的房子...
     Area  No. of Rooms  No. of Bathrooms Location  Miles (dist. between school and house)  Rent Price per Month  Sell Price
4    2138             1                 1   Suburb                                      10                  9923    50273384
18   1455             2                 1   Suburb                                      30                  8222    25214047
404  2765             3                 1   Suburb                                      32                  7991    25596398
486  1890             1                 1   Suburb                                      46                  7907    12427095
529  2714             1                 1   Suburb                                      30                  6882     6880096
669  2645             1                 1   Suburb                                      44                  7813    54025351
689  2600             1                 1   Suburb                                      46                  8093    23353731
954  2943             1                 1   Suburb                                      37                  7920    49070149

[INFO] Rural 符合租房條件的房子...
     Area  No. of Rooms  No. of Bathrooms Location  Miles (dist. between school and house)  Rent Price per Month  Sell Price
345  2852             3                 1    Rural                                      37                  6278    48500282
646  2074             1                 1    Rural                                      22                  9849    35731980
781  1993             3                 1    Rural                                      32                  6294    24235532
888  2636             2                 1    Rural                                      33                  8773     8112368
991   917             3                 1    Rural                                      26                  9893    18433238
```

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?

我會選擇買City Center 價格4881萬 距離學校10 miles 面積2574 三房的房子。因為它符合我的篩選條件離學校近、兩房以上、面積大、價格不高。

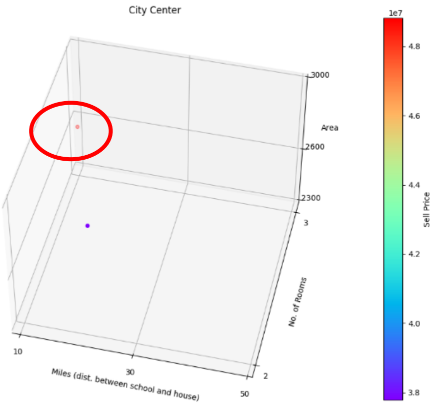
1. 先打印出City Center的describe() 看篩選的範圍大概要設在哪：



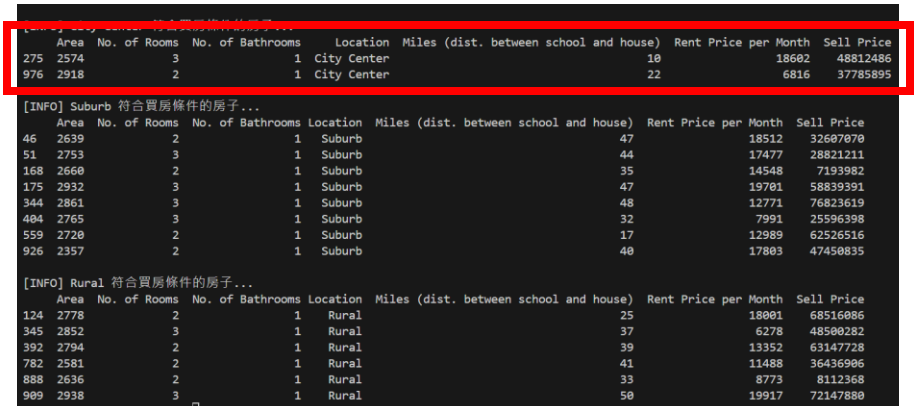
```
[INFO] City Center describe...
              Area  No. of Rooms  No. of Bathrooms  Miles (dist. between school and house)  Rent Price per Month    Sell Price
count   352.000000    352.000000             352.0                              352.000000            352.000000  3.520000e+02
mean   1743.187500      1.931818               1.0                              261.428977          13272.215909  4.189646e+07
std     688.234671      0.810128               0.0                              140.708648           4069.382362  2.155917e+07
min     505.000000      1.000000               1.0                               10.000000           6062.000000  6.131936e+06
25%    1166.750000      1.000000               1.0                              136.750000           9661.500000  2.368423e+07
50%    1714.500000      2.000000               1.0                              270.000000          13276.500000  4.287159e+07
75%    2324.750000      3.000000               1.0                              385.250000          17024.500000  6.095799e+07
max    2992.000000      3.000000               1.0                              498.000000          19907.000000  7.998578e+07
```

因為錢不是問題，所以篩選條件只考慮學校距離、房屋面積、房間數量：

X 軸：學校距離 Y 軸：房間數量 Z 軸：房屋面積 圖上的散點：點的顏色來區分房價



2. 決定買 City Center 價格 4881 萬 距離學校 10 miles 面積 2574 三房的房子：



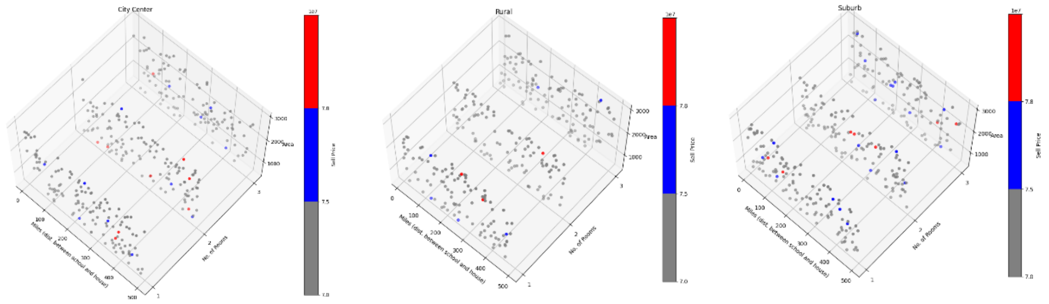
```
[INFO] City Center 符合買房條件的房子...
     Area  No. of Rooms  No. of Bathrooms     Location  Miles (dist. between school and house)  Rent Price per Month  Sell Price
275  2574             3                 1  City Center                                      10                 18602    48812486
976  2918             2                 1  City Center                                      22                  6816    37785895

[INFO] Suburb 符合買房條件的房子...
     Area  No. of Rooms  No. of Bathrooms Location  Miles (dist. between school and house)  Rent Price per Month  Sell Price
46   2639             2                 1   Suburb                                      47                 18512    32607070
51   2753             3                 1   Suburb                                      44                 17477    28821211
168  2660             2                 1   Suburb                                      35                 14548     7193982
175  2932             3                 1   Suburb                                      47                 19701    58839391
344  2861             3                 1   Suburb                                      48                 12771    76823619
404  2765             3                 1   Suburb                                      32                  7991    25596398
559  2720             2                 1   Suburb                                      17                 12989    62526516
926  2357             2                 1   Suburb                                      40                 17803    47450835

[INFO] Rural 符合買房條件的房子...
     Area  No. of Rooms  No. of Bathrooms Location  Miles (dist. between school and house)  Rent Price per Month  Sell Price
124  2778             2                 1    Rural                                      25                 18001    68516086
345  2852             3                 1    Rural                                      37                  6278    48500282
392  2794             2                 1    Rural                                      39                 13352    63147728
782  2581             2                 1    Rural                                      41                 11488    36436906
888  2636             2                 1    Rural                                      33                  8773     8112368
909  2938             3                 1    Rural                                      50                 19917    72147880
```

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

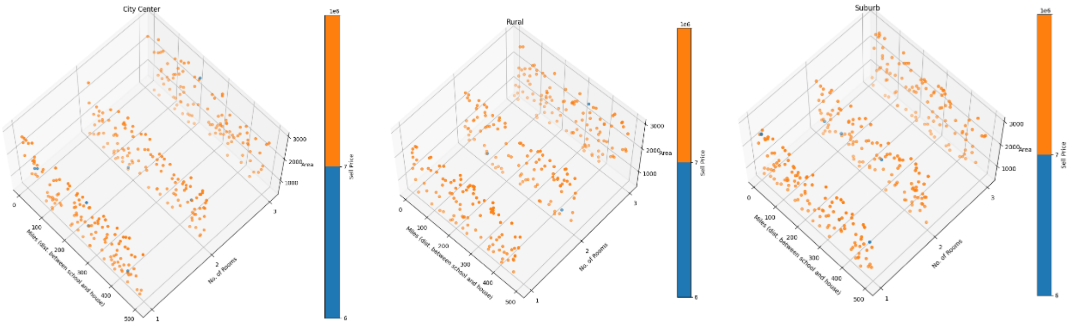
有幾間售價特別高跟特別低的房子，但從三維的散點圖來看，不管是面積、房間數量、地點、學校的距離都看不出造成高低價的關係。

X 軸：學校距離 Y 軸：房間數量 Z 軸：房屋面積 圖上的散點：點的顏色來區分房價

下面三張圖為 City Center、Rural、Suburb 的高房價分布圖



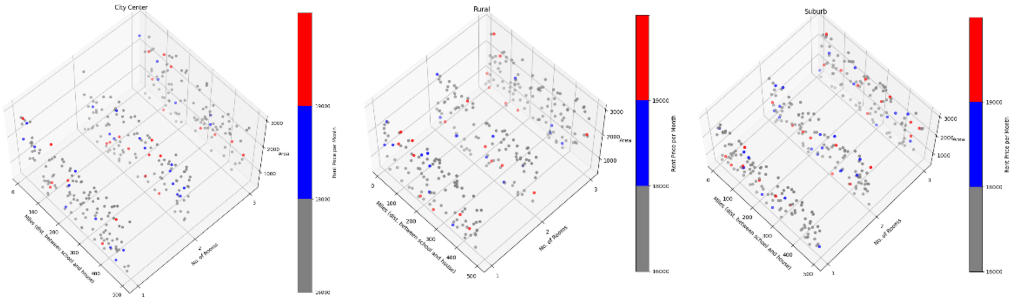
下面三張圖為 City Center、Rural、Suburb 的低房價分布圖



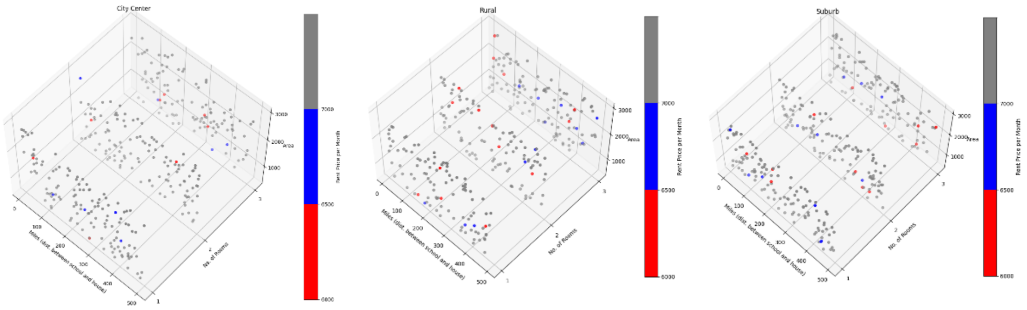
有幾間房租特別高跟特別低的房子，但從三維的散點圖來看，不管是面積、房間數量、地點、學校的距離都看不出造成高低價的關係

X 軸：學校距離 Y 軸：房間數量 Z 軸：房屋面積 圖上的散點：點的顏色來區分房價

下面三張圖為 City Center、Rural、Suburb 的高房租分布圖



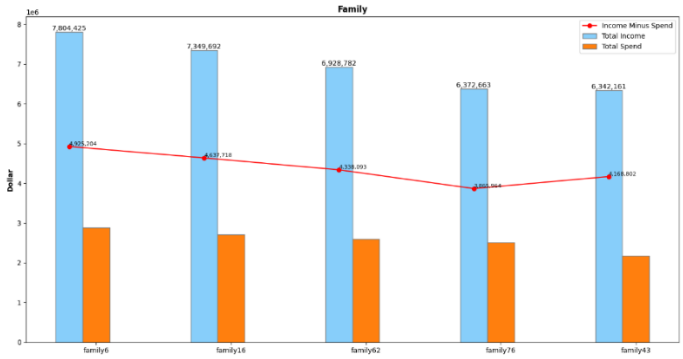
下面三張圖為 City Center、Rural、Suburb 的低房租分布圖



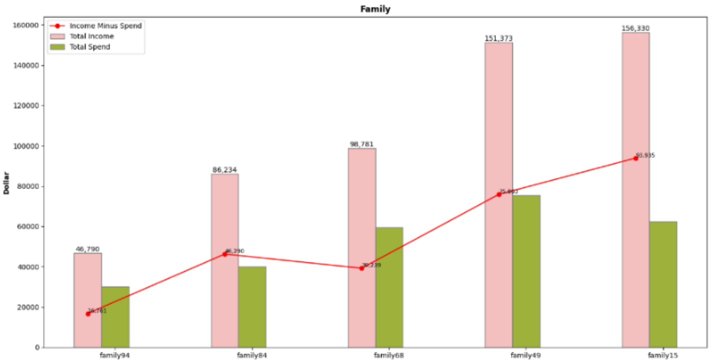
Dataset 2: family_data.csv

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

收入最多的家庭：family6

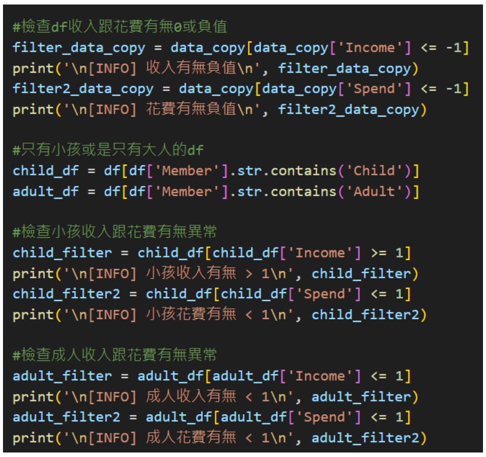


收入最少的家庭：family94

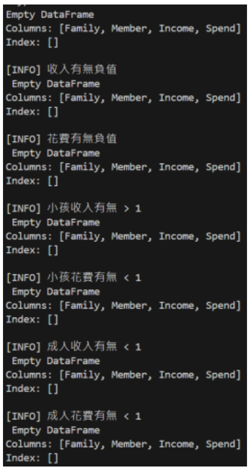


以下為篩選過程

1.　檢查大人小孩的收入跟花費有無異常：　無異常

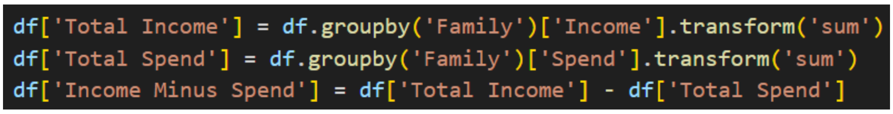


```
#檢查df收入跟花費有無0或負值
filter_data_copy = data_copy[data_copy['Income'] <= -1]
print('\n[INFO] 收入有無負值\n', filter_data_copy)
filter2_data_copy = data_copy[data_copy['Spend'] <= -1]
print('\n[INFO] 花費有無負值\n', filter2_data_copy)

#只有小孩或是只有大人的df
child_df = df[df['Member'].str.contains('Child')]
adult_df = df[df['Member'].str.contains('Adult')]

#檢查小孩收入跟花費有無異常
child_filter = child_df[child_df['Income'] >= 1]
print('\n[INFO] 小孩收入有無 > 1\n', child_filter)
child_filter2 = child_df[child_df['Spend'] <= 1]
print('\n[INFO] 小孩花費有無 < 1\n', child_filter2)

#檢查成人收入跟花費有無異常
adult_filter = adult_df[adult_df['Income'] <= 1]
print('\n[INFO] 成人收入有無 < 1\n', adult_filter)
adult_filter2 = adult_df[adult_df['Spend'] <= 1]
print('\n[INFO] 成人花費有無 < 1\n', adult_filter2)
```



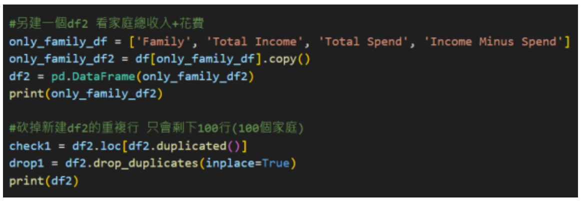
```
Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []

[INFO] 收入有無負值
 Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []

[INFO] 花費有無負值
 Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []

[INFO] 小孩收入有無 > 1
 Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []

[INFO] 小孩花費有無 < 1
 Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []

[INFO] 成人收入有無 < 1
 Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []

[INFO] 成人花費有無 < 1
 Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []
```

2. 新增三個欄位： 家庭收入加總、家庭花費加總、家庭收入減掉花費

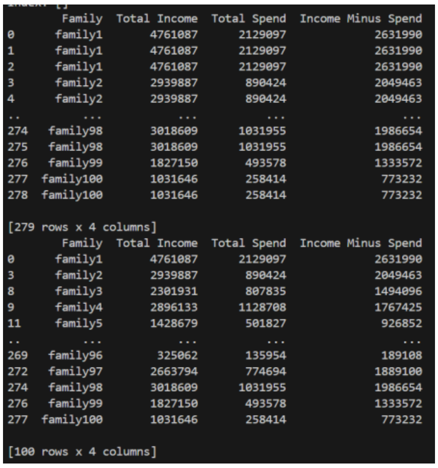


```
df['Total Income'] = df.groupby('Family')['Income'].transform('sum')
df['Total Spend'] = df.groupby('Family')['Spend'].transform('sum')
df['Income Minus Spend'] = df['Total Income'] - df['Total Spend']
```

3. 新建一個dataframe2：只有家庭跟上面新增的三欄 刪除重複值 得到了100個家庭的收入跟支出

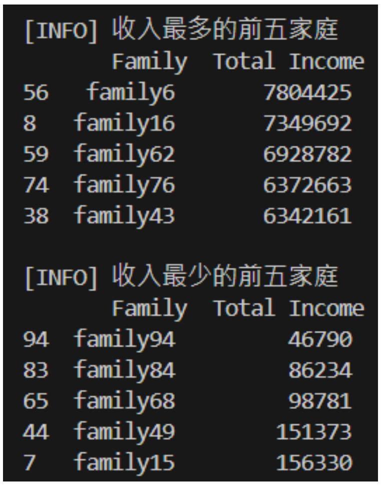


```
#另建一個df2 看家庭總收入+花費
only_family_df = ['Family', 'Total Income', 'Total Spend', 'Income Minus Spend']
only_family_df2 = df[only_family_df].copy()
df2 = pd.DataFrame(only_family_df2)
print(only_family_df2)

#砍掉新建df2的重複行 只會剩下100行(100個家庭)
check1 = df2.loc[df2.duplicated()]
drop1 = df2.drop_duplicates(inplace=True)
print(df2)
```



```
       Family  Total Income  Total Spend  Income Minus Spend
0     family1       4761087      2129097             2631990
1     family1       4761087      2129097             2631990
2     family1       4761087      2129097             2631990
3     family2       2939887       890424             2049463
4     family2       2939887       890424             2049463
..        ...           ...          ...                 ...
274  family98       3018609      1031955             1986654
275  family98       3018609      1031955             1986654
276  family99       1827150       493578             1333572
277 family100       1031646       258414              773232
278 family100       1031646       258414              773232

[279 rows x 4 columns]
       Family  Total Income  Total Spend  Income Minus Spend
0     family1       4761087      2129097             2631990
3     family2       2939887       890424             2049463
8     family3       2301931       807835             1494096
9     family4       2896133      1128708             1767425
11    family5       1428679       501827              926852
..        ...           ...          ...                 ...
269  family96        325062       135954              189108
272  family97       2663794       774694             1889100
274  family98       3018609      1031955             1986654
276  family99       1827150       493578             1333572
277 family100       1031646       258414              773232

[100 rows x 4 columns]
```

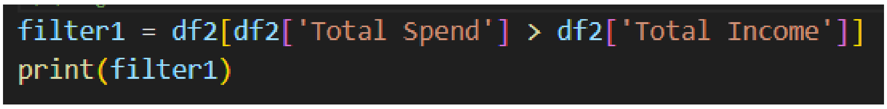
4. 找到新建dataframe2裡最高跟最低的家庭收入 畫出最上面的兩張bar+折線圖



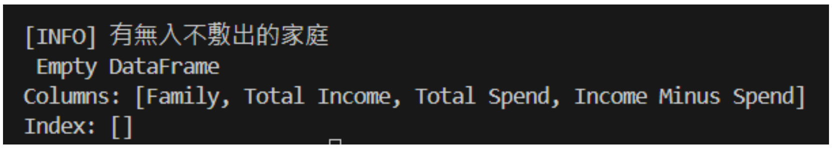
```
[INFO] 收入最多的前五家庭
      Family  Total Income
56   family6       7804425
8   family16       7349692
59  family62       6928782
74  family76       6372663
38  family43       6342161

[INFO] 收入最少的前五家庭
      Family  Total Income
94  family94         46790
83  family84         86234
65  family68         98781
44  family49        151373
7   family15        156330
```

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

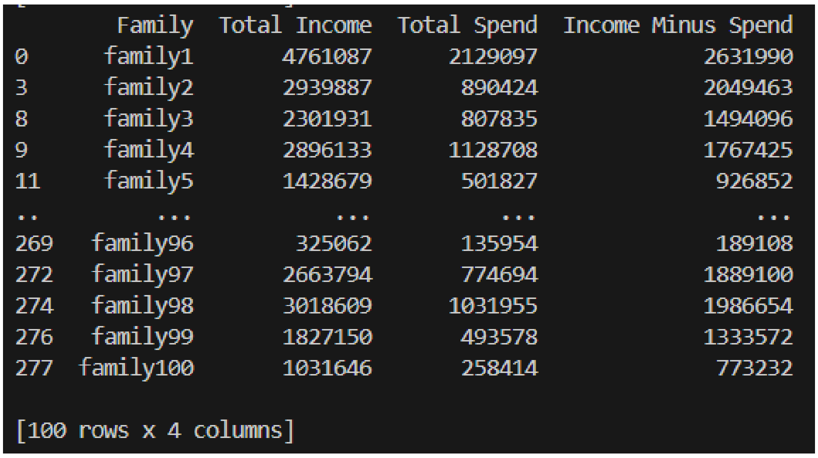
我的答案是沒有。我用Q1提到的dataframe2來尋找花費大於支出的家庭，得到的回應是Empty DataFrame。



```
filter1 = df2[df2['Total Spend'] > df2['Total Income']]
print(filter1)
```



```
[INFO] 有無入不敷出的家庭
 Empty DataFrame
Columns: [Family, Total Income, Total Spend, Income Minus Spend]
Index: []
```



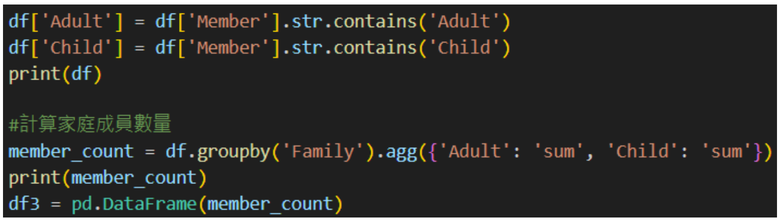
```
       Family  Total Income  Total Spend  Income Minus Spend
0     family1       4761087      2129097             2631990
3     family2       2939887       890424             2049463
8     family3       2301931       807835             1494096
9     family4       2896133      1128708             1767425
11    family5       1428679       501827              926852
..        ...           ...          ...                 ...
269  family96        325062       135954              189108
272  family97       2663794       774694             1889100
274  family98       3018609      1031955             1986654
276  family99       1827150       493578             1333572
277 family100       1031646       258414              773232

[100 rows x 4 columns]
```

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

40個單親家庭 35個無子女家庭

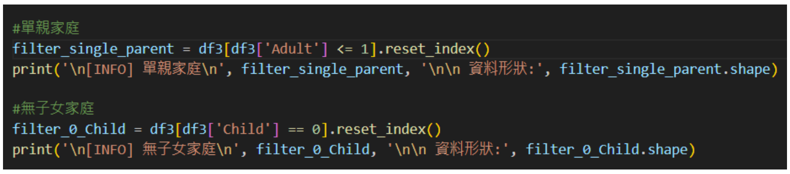
以下為篩選程式碼：

1. 新增每個家庭的大人跟小孩欄位回傳布林值
2. 計算成員數量



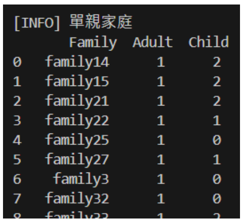
```
df['Adult'] = df['Member'].str.contains('Adult')
df['Child'] = df['Member'].str.contains('Child')
print(df)

#計算家庭成員數量
member_count = df.groupby('Family').agg({'Adult': 'sum', 'Child': 'sum'})
print(member_count)
df3 = pd.DataFrame(member_count)
```

3. 設定篩選條件

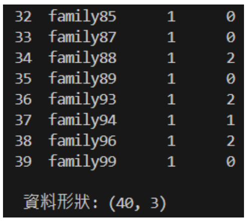


```
#單親家庭
filter_single_parent = df3[df3['Adult'] <= 1].reset_index()
print('\n[INFO] 單親家庭\n', filter_single_parent, '\n\n 資料形狀:', filter_single_parent.shape)

#無子女家庭
filter_0_Child = df3[df3['Child'] == 0].reset_index()
print('\n[INFO] 無子女家庭\n', filter_0_Child, '\n\n 資料形狀:', filter_0_Child.shape)
```

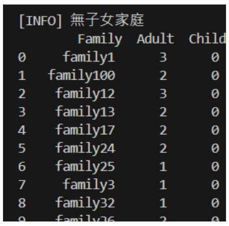
4. 回傳篩選的資料形狀



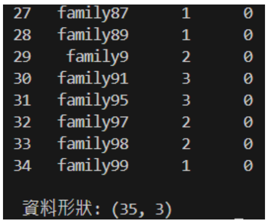
```
[INFO] 單親家庭
      Family  Adult  Child
0   family14      1      2
1   family15      1      2
2   family21      1      2
3   family22      1      1
4   family25      1      0
5   family27      1      1
6    family3      1      0
7   family32      1      0
```



```
32  family85      1      0
33  family87      1      0
34  family88      1      2
35  family89      1      0
36  family93      1      2
37  family94      1      1
38  family96      1      2
39  family99      1      0

 資料形狀: (40, 3)
```



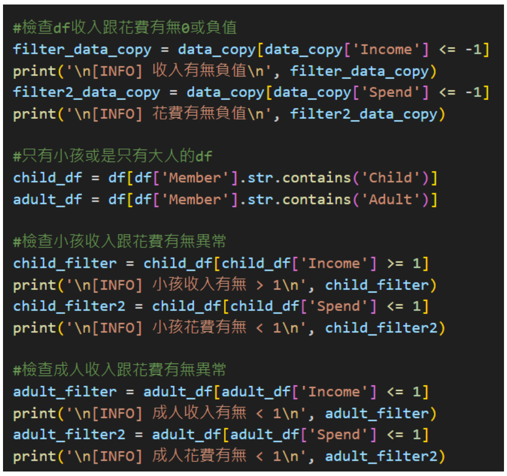
```
[INFO] 無子女家庭
       Family  Adult  Child
0     family1      3      0
1   family100      2      0
2    family12      3      0
3    family13      2      0
4    family17      2      0
5    family24      2      0
6    family25      1      0
7     family3      1      0
8    family32      1      0
```



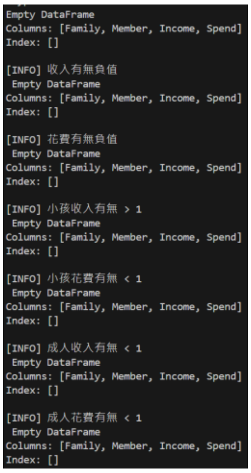
```
27  family87      1      0
28  family89      1      0
29   family9      2      0
30  family91      3      0
31  family95      3      0
32  family97      2      0
33  family98      2      0
34  family99      1      0

 資料形狀: (35, 3)
```

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

我的答案是沒有，處理資料前有先檢查有無缺失或重複值，也有針對大人小孩來過濾看看有沒有不合理的地方。



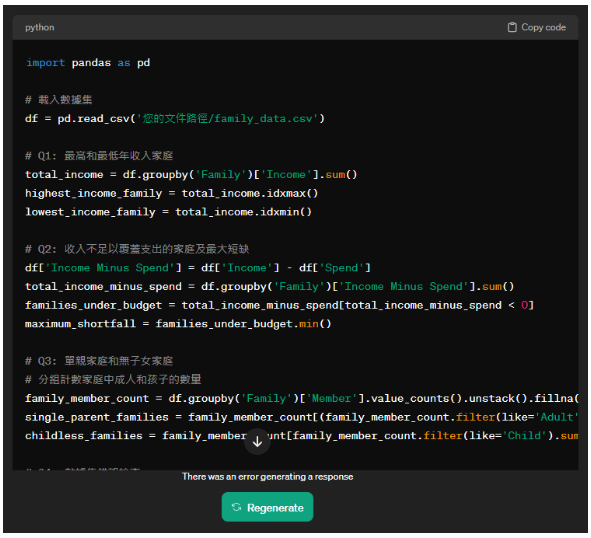
```
#檢查df收入跟花費有無0或負值
filter_data_copy = data_copy[data_copy['Income'] <= -1]
print('\n[INFO] 收入有無負值\n', filter_data_copy)
filter2_data_copy = data_copy[data_copy['Spend'] <= -1]
print('\n[INFO] 花費有無負值\n', filter2_data_copy)

#只有小孩或是只有大人的df
child_df = df[df['Member'].str.contains('Child')]
adult_df = df[df['Member'].str.contains('Adult')]

#檢查小孩收入跟花費有無異常
child_filter = child_df[child_df['Income'] >= 1]
print('\n[INFO] 小孩收入有無 > 1\n', child_filter)
child_filter2 = child_df[child_df['Spend'] <= 1]
print('\n[INFO] 小孩花費有無 < 1\n', child_filter2)

#檢查成人收入跟花費有無異常
adult_filter = adult_df[adult_df['Income'] <= 1]
print('\n[INFO] 成人收入有無 < 1\n', adult_filter)
adult_filter2 = adult_df[adult_df['Spend'] <= 1]
print('\n[INFO] 成人花費有無 < 1\n', adult_filter2)
```



```
Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []

[INFO] 收入有無負值
 Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []

[INFO] 花費有無負值
 Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []

[INFO] 小孩收入有無 > 1
 Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []

[INFO] 小孩花費有無 < 1
 Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []

[INFO] 成人收入有無 < 1
 Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []

[INFO] 成人花費有無 < 1
 Empty DataFrame
Columns: [Family, Member, Income, Spend]
Index: []
```

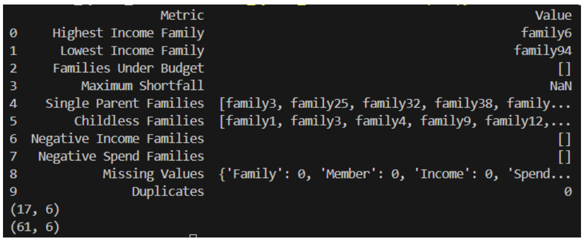
Q5. Can ChatGPT or Bing assist with the aforementioned four questions? If so, to what extent? How do you issue commands to the AI tool? If not, why not?

我把 dataset 上傳到 ChatGPT 直接問他 Q1~Q4 以下是他給我的程式碼跟執行後的結果



```
import pandas as pd

# 載入數據集
df = pd.read_csv('您的文件路徑/family_data.csv')

# Q1: 最高和最低年收入家庭
total_income = df.groupby('Family')['Income'].sum()
highest_income_family = total_income.idxmax()
lowest_income_family = total_income.idxmin()

# Q2: 收入不足以覆蓋支出的家庭及最大短缺
df['Income Minus Spend'] = df['Income'] - df['Spend']
total_income_minus_spend = df.groupby('Family')['Income Minus Spend'].sum()
families_under_budget = total_income_minus_spend[total_income_minus_spend < 0]
maximum_shortfall = families_under_budget.min()

# Q3: 單親家庭和無子女家庭
# 分組計數家庭中成人和孩子的數量
family_member_count = df.groupby('Family')['Member'].value_counts().unstack().fillna(
single_parent_families = family_member_count[(family_member_count.filter(like='Adult'
childless_families = family_member   unt[family_member_count.filter(like='Child').sum
```

There was an error generating a response



```
                    Metric                                              Value
0    Highest Income Family                                            family6
1     Lowest Income Family                                           family94
2    Families Under Budget                                                 []
3        Maximum Shortfall                                                NaN
4   Single Parent Families  [family3, family25, family32, family38, family...
5       Childless Families  [family1, family3, family4, family9, family12,...
6 Negative Income Families                                                 []
7  Negative Spend Families                                                 []
8           Missing Values  {'Family': 0, 'Member': 0, 'Income': 0, 'Spend...
9               Duplicates                                                  0
(17, 6)
(61, 6)
```