Week 3: Recall the data from an excel file which was used for intermediate saving of the "df_cleaned" data frame

The data frame contains all relevant (and modified) data:

```
age  sex  householdincome  howoftenwine  noofwines  \
0       34    2               12            10          1
1       84    2                7             6          1
2       29    2               13            10          1
3       68    2                6             5          1
4       54    2               11             9          1
...    ...  ...              ...           ...        ...
14556   18    2                1             9          1
14557   18    1                1            10          1
14558   51    1                6             6          1
14559   21    1                1            10          2
14560   18    2                1            10          1

                  wine_frequency              wine_amount  \
0      1 or 2 times in the last year     One glass/ container
1                2 to 3 times a month     One glass/ container
2      1 or 2 times in the last year     One glass/ container
3                        Once a week     One glass/ container
4      3 to 6 times in the last year     One glass/ container
...                              ...                      ...
14556  3 to 6 times in the last year     One glass/ container
14557  1 or 2 times in the last year     One glass/ container
14558            2 to 3 times a month     One glass/ container
14559  1 or 2 times in the last year  Two glasses/ containers
14560  1 or 2 times in the last year     One glass/ container

      income_category
    0      $50,000 to $59,999
    1      $20,000 to $24,999
    2      $60,000 to $69,999
    3      $15,000 to $19,999
    4      $40,000 to $49,999
    ...                    ...
    14556    Less than $5,000
14557    Less than $5,000
    14558  $15,000 to $19,999
    14559    Less than $5,000
14560    Less than $5,000

    [14561 rows x 8 columns]
```

As stated in week 2, the "age" data is not binned yet, which makes the processing and plotting of the data more difficult. Hence, binning is done first.

In [9]:

```
# Create bins for Age (note that the age was ranging from 18 to 96)
bins = [18, 24, 34, 44, 54, 64, 74, 84, 94, 104]  # Define the bins
labels = ['18-24', '25-34', '35-44', '45-54', '55-64', '65-74', '75-84', '85-94', '95

# Create a new column 'age_group' with the binned age data
df_cleaned['age_group'] = pd.cut(df_cleaned['age'], bins=bins, labels=labels, right=F

# Display the updated data frame
print(df_cleaned)
```

```
       age  sex  householdincome  howoftenwine  noofwines  \
0       34    2               12            10          1
1       84    2                7             6          1
2       29    2               13            10          1
3       68    2                6             5          1
4       54    2               11             9          1
...    ...  ...              ...           ...        ...
14556   18    2                1             9          1
14557   18    1                1            10          1
14558   51    1                6             6          1
14559   21    1                1            10          2
14560   18    2                1            10          1

                      wine_frequency              wine_amount  \
0      1 or 2 times in the last year     One glass/ container
1                2 to 3 times a month     One glass/ container
2      1 or 2 times in the last year     One glass/ container
3                         Once a week     One glass/ container
4      3 to 6 times in the last year     One glass/ container
...                              ...                      ...
14556  3 to 6 times in the last year     One glass/ container
14557  1 or 2 times in the last year     One glass/ container
14558            2 to 3 times a month     One glass/ container
14559  1 or 2 times in the last year  Two glasses/ containers
14560  1 or 2 times in the last year     One glass/ container

          income_category age_group
0      $50,000 to $59,999     35-44
1      $20,000 to $24,999     85-94
2      $60,000 to $69,999     25-34
3      $15,000 to $19,999     65-74
4      $40,000 to $49,999     55-64
...                   ...       ...
14556    Less than $5,000     18-24
14557    Less than $5,000     18-24
14558  $15,000 to $19,999     45-54
14559    Less than $5,000     18-24
14560    Less than $5,000     18-24

[14561 rows x 9 columns]
```

Since the initial columns "age", "householdincome", "howoftenwine" and "noofwines" are not needed anymore, they are removed from the data frame. Further, the columns are rearranged in a meaningful order.

In [11]:

```
# Drop the columns "age", "householdincome", "howoftenwine" and "noofwines"
df_cleaned = df_cleaned.drop(columns=['age', 'householdincome', 'howoftenwine', 'noof

# Rearrange the columns
df_cleaned = df_cleaned[['age_group', 'sex', 'wine_frequency', 'wine_amount', 'income

# Display the updated data frame
print(df_cleaned)
```

```
      age_group  sex                 wine_frequency              wine_amount  \
0         35-44    2  1 or 2 times in the last year     One glass/ container
1         85-94    2           2 to 3 times a month     One glass/ container
2         25-34    2  1 or 2 times in the last year     One glass/ container
3         65-74    2                    Once a week     One glass/ container
4         55-64    2  3 to 6 times in the last year     One glass/ container
...         ...  ...                             ...                      ...
14556     18-24    2  3 to 6 times in the last year     One glass/ container
14557     18-24    1  1 or 2 times in the last year     One glass/ container
14558     45-54    1           2 to 3 times a month     One glass/ container
14559     18-24    1  1 or 2 times in the last year  Two glasses/ containers
14560     18-24    2  1 or 2 times in the last year     One glass/ container

          income_category
0      $50,000 to $59,999
1      $20,000 to $24,999
2      $60,000 to $69,999
3      $15,000 to $19,999
4      $40,000 to $49,999
...                   ...
14556    Less than $5,000
14557    Less than $5,000
14558  $15,000 to $19,999
14559    Less than $5,000
14560    Less than $5,000

[14561 rows x 5 columns]
```

The final result is now:

```
   age_group  sex                 wine_frequency
   wine_amount  \
   0         35-44    2  1 or 2 times in the last year     One glass/
   container
   1         85-94    2           2 to 3 times a month     One glass/
   container
   2         25-34    2  1 or 2 times in the last year     One glass/
   container
   3         65-74    2                    Once a week     One glass/
   container
   4         55-64    2  3 to 6 times in the last year     One glass/
   container
   ...         ...  ...                             ...
   ...
   14556     18-24    2  3 to 6 times in the last year     One glass/
   container
```

```
14557     18-24    1  1 or 2 times in the last year     One glass/
container
14558     45-54    1           2 to 3 times a month     One glass/
container
14559     18-24    1  1 or 2 times in the last year  Two glasses/
containers
14560     18-24    2  1 or 2 times in the last year     One glass/
container

      income_category
0      $50,000 to $59,999
1      $20,000 to $24,999
2      $60,000 to $69,999
3      $15,000 to $19,999
4      $40,000 to $49,999
...                   ...
14556    Less than $5,000
```

14557 Less than $5,000 1455815,000 to $19,999 14559 $Less than$5,000

14560 Less than $5,000

```
[14561 rows x 5 columns]
```

From my point of view, the data is now prepared well for further processing.