In [16]:

```
import pandas as pd
import numpy as np
import scipy.stats
import seaborn as sns
import matplotlib.pyplot as plt

# Reading the data which was saved to an excel sheet after pre-processing
df_cleaned = pd.read_excel('finally_clean_data_for_plotting.xlsx')

# Remove the 99 category (="Unknown") from the data, since it does not benefit the ar
df_cleaned = df_cleaned[~(df_cleaned['noofwines'] == 99)]
df_cleaned = df_cleaned[~(df_cleaned['howoftenwine'] == 99)]

print(df_cleaned)
```

Recalling the dataset, that I have used for the last assignments:

```
         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 $5,0001455815$,000 to $19,99914559 Less than 5$,000
14560 Less than $5,000

[14561 rows x 8 columns]

I would like to know how the relationships between age and wine consumption vary across different income categories. Thus, I chose the income as a moderator for my dataset:

In [22]:

```
# Performing the correlation without any moderation
print (scipy.stats.pearsonr(df_cleaned['age'], df_cleaned['noofwines']))

# Define 3 different income groups, as in the given example (refer to codebook)
# 1: Less than $5,000 to $24,999
# 2: $25,000 to $69,999
# 3: $70,000 to $200,000 or more
def incomegrp (row):
   if row['householdincome'] <= 7:
      return 1
   elif row['householdincome'] <= 14:
      return 2
   elif row['householdincome'] > 14:
      return 3

df_cleaned['incomegrp'] = df_cleaned.apply (lambda row: incomegrp (row),axis=1)

chk1 = df_cleaned['incomegrp'].value_counts(sort=False, dropna=False)
print(chk1)

sub1 = df_cleaned[(df_cleaned['incomegrp'] == 1)]
sub2 = df_cleaned[(df_cleaned['incomegrp'] == 2)]
sub3 = df_cleaned[(df_cleaned['incomegrp'] == 3)]

print ('       ')
print ('Association between age and wine drinking amount for LOW income countries')
print (scipy.stats.pearsonr(sub1['age'], sub1['noofwines']))
print ('       ')
print ('Association between age and wine drinking amount for MIDDLE income countries
print (scipy.stats.pearsonr(sub2['age'], sub2['noofwines']))
print ('       ')
print ('Association between age and wine drinking amount for HIGH income countries')
print (scipy.stats.pearsonr(sub3['age'], sub3['noofwines']))
```

```
PearsonRResult(statistic=-0.15769467412496058, pvalue=1.6732982798117926e-81)
incomegrp
2    7796
1    3137
3    3589
Name: count, dtype: int64

Association between age and wine drinking amount for LOW income countries
PearsonRResult(statistic=-0.20243650427576287, pvalue=2.2845360210949472e-30)

Association between age and wine drinking amount for MIDDLE income countries
PearsonRResult(statistic=-0.1488408911690387, pvalue=7.337846604962096e-40)

Association between age and wine drinking amount for HIGH income countries
```

```
PearsonRResult(statistic=-0.11098519536420046, pvalue=2.606939038381185e-11)
```

Interpretation of results:

```
1. Association between age and wine drinking amount for LOW income
countries
PearsonRResult(statistic=-0.20243650427576287,
pvalue=2.2845360210949472e-30)
```

There is a statistically significant negative correlation between age and the number of wines consumed in the low income group. This suggests that as age increases, the number of wines consumed tends to decrease.

```
2. Association between age and wine drinking amount for MIDDLE income
countries
PearsonRResult(statistic=-0.1488408911690387,
pvalue=7.337846604962096e-40)
```

Similarly, there is a statistically significant negative correlation between age and wine consumption in the middle-income group, although the strength of the correlation is weaker than in the low-income group.

```
3. Association between age and wine drinking amount for HIGH income
countries
PearsonRResult(statistic=-0.11098519536420046,
pvalue=2.606939038381185e-11)
```

There is also a statistically significant negative correlation in the high-income group, but again, it is the weakest among the three groups.

Summary: The introduction of a moderator does not change my results, gained from the first approach using the un-moderated correlation coefficient (see last assignment).