In [ ]:

```
import pandas as pd
import numpy as np
import scipy
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)
```

```
          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]

Now let's prepare the data for creating a correlation coefficient. I decided to select again the variables relevant for my hypotheses.

In [ ]:
```
print ('Association between age and wine drinking amount:')
print (scipy.stats.pearsonr(df_cleaned['age'], df_cleaned['noofwines']))
print ()
print ()
print ('Association between household income and  wine drinking amount:')
print (scipy.stats.pearsonr(df_cleaned['householdincome'], df_cleaned['noofwines']))
print ()
print ()
print ('Association between age and  household income:')
print (scipy.stats.pearsonr(df_cleaned['age'], df_cleaned['householdincome']))
```

Result:

```
Association between age and wine drinking amount:
PearsonRResult(statistic=-0.15769467412496058,
pvalue=1.6732982798117926e-81)


Association between household income and wine drinking amount:
PearsonRResult(statistic=0.007533398347107097,
pvalue=0.3640029559239077)


Association between age and household income:
PearsonRResult(statistic=-0.04895724698791094,
pvalue=3.5752358021344056e-09)
```

Interpretation of results:

Correlation between age and noofwines:

The correlation coefficient is -0.1577. This indicates a weak negative correlation between age and the number of wines consumed. This means that as age increases, drinking amount tends to decrease.

Correlation between householdincome and noofwines:

The correlation coefficient is 0.0075. This indicates a very weak positive correlation between household income and the number of wines consumed. Similar to the previous correlations, this suggests that there is almost no relationship between these two variables.

Correlation between age and householdincome:

The correlation coefficient is -0.04896. This indicates a very weak negative correlation between age and household income. This means that as age increases, household income tends to decrease slightly.

Since I did not figure out any satisfying correlation, I would also like to include the drinking frequency in my analysis, to see, whether there is a correlation to either age, amount of wine drank or household income.

In [ ]:

```
print ('Association between age and wine drinking frequency:')
print (scipy.stats.pearsonr(df_cleaned['age'], df_cleaned['howoftenwine']))
print ()
print ()
print ('Association between household income and wine drinking frequency:')
print (scipy.stats.pearsonr(df_cleaned['householdincome'], df_cleaned['howoftenwine']
print ()
print ()
print ('Association between wine drinking amount and household income:')
print (scipy.stats.pearsonr(df_cleaned['noofwines'], df_cleaned['howoftenwine']))
```

Result:

```
Association between age and wine drinking frequency:
PearsonRResult(statistic=-0.1845837406173188,
pvalue=1.801305078469256e-111)


Association between household income and wine drinking frequency:
PearsonRResult(statistic=-0.1471527757744252,
pvalue=4.240279040114272e-71)


Association between wine drinking amount and household income:
PearsonRResult(statistic=-0.15921606183282871,
pvalue=4.5717587860949234e-83)
```

Interpretation of the results:

Correlation between age and wine drinking frequency: The correlation coefficient is -0.1846. This indicates a weak negative correlation, and thus, a slight decrease of wine drinking frequency with increasing age.

Correlation between household income and wine drinking frequency: The correlation coefficient is -0.1472. This indicates a weak negative correlation, and thus, a slight decrease of wine drinking frequency with increasing household income.

Correlation between wine drinking amount and household income: The correlation coefficient is -0.1592. This indicates a weak negative correlation, and thus, a slight decrease of wine drinking frequency with increasing wine drinking amount.

Overall, the correlation of the variables to wine drinking frequency seems to be stronger, than compared to the other ones.