

In [2]:

```
import pandas as pd
import matplotlib.pyplot as plt

df_cleaned = pd.read_excel('finally_clean_data_for_further_processing.xlsx')

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

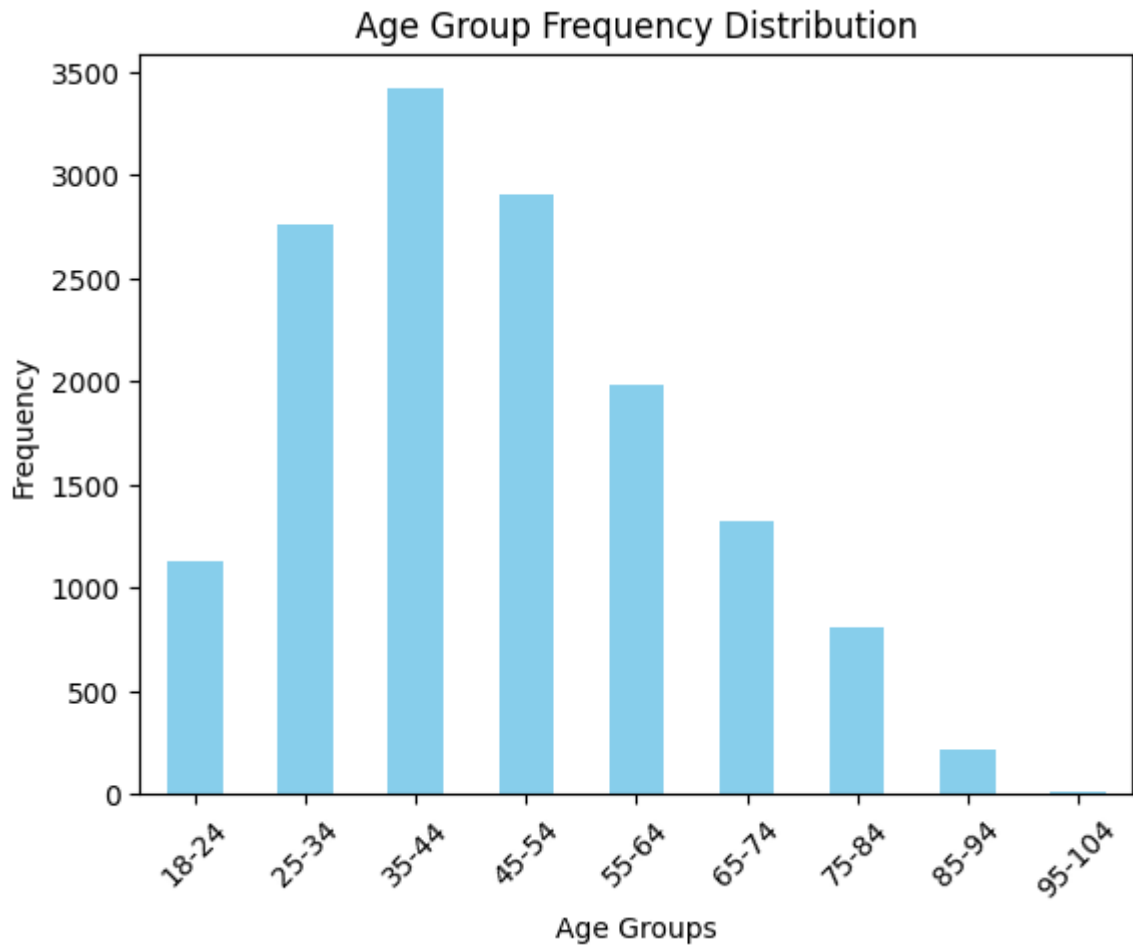
[14561 rows x 6 columns]

Generating the univariate graph to illustrate the distributions of week 2. Please note, that "sex" was not initially chosen to be part of my hypotheses. Hence, the "sex" attribute is not used any further, but maybe relevant for future analyses of the drinking behaviour.

In [2]:

```
# Plotting the Age Group Frequency Distribution
labels = ['18-24', '25-34', '35-44', '45-54', '55-64', '65-74', '75-84', '85-94', '95-104']
age_group = df_cleaned['age_group'].value_counts().reindex(labels, fill_value=0).sort_index()

age_group.plot(kind='bar', color='skyblue')
plt.title('Age Group Frequency Distribution')
plt.xlabel('Age Groups')
plt.ylabel('Frequency')
plt.xticks(rotation=45)
plt.show()
```



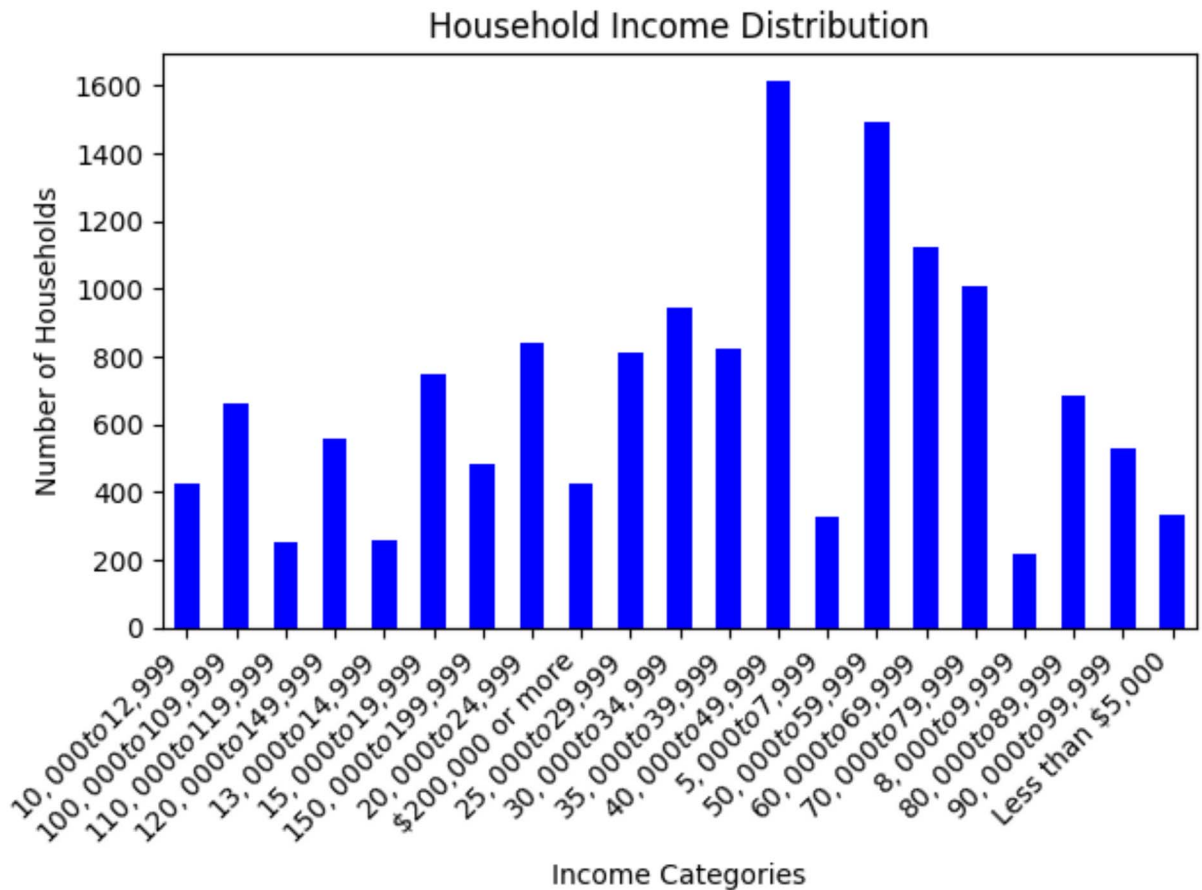
The distribution of age groups has a peak at 35-44 years. Second and third most age groups are 45-54 and 25-34. The other age groups are less represented in the sample.

In [3]:

```
# Plotting the Income Category Frequency Distribution
income_labels = ['Less than $5,000',
                 '$5,000 to $7,999',
                 '$8,000 to $9,999',
                 '$10,000 to $12,999',
                 '$13,000 to $14,999',
                 '$15,000 to $19,999',
                 '$20,000 to $24,999',
                 '$25,000 to $29,999',
                 '$30,000 to $34,999',
                 '$35,000 to $39,999',
                 '$40,000 to $49,999',
                 '$50,000 to $59,999',
                 '$60,000 to $69,999',
                 '$70,000 to $79,999',
                 '$80,000 to $89,999',
                 '$90,000 to $99,999',
                 '$100,000 to $109,999',
                 '$110,000 to $119,999',
                 '$120,000 to $149,999',
                 '$150,000 to $199,999',
                 '$200,000 or more']

income_data = df_cleaned['income_category'].value_counts().reindex(income_labels, fill_value=0)

income_data.plot(kind='bar', color='blue')
plt.title('Household Income Distribution')
plt.xlabel('Income Categories')
plt.ylabel('Number of Households')
plt.xticks(rotation=45, ha='right')
plt.tight_layout() # Adjust Layout to make room for x-axis Labels
plt.show()
```

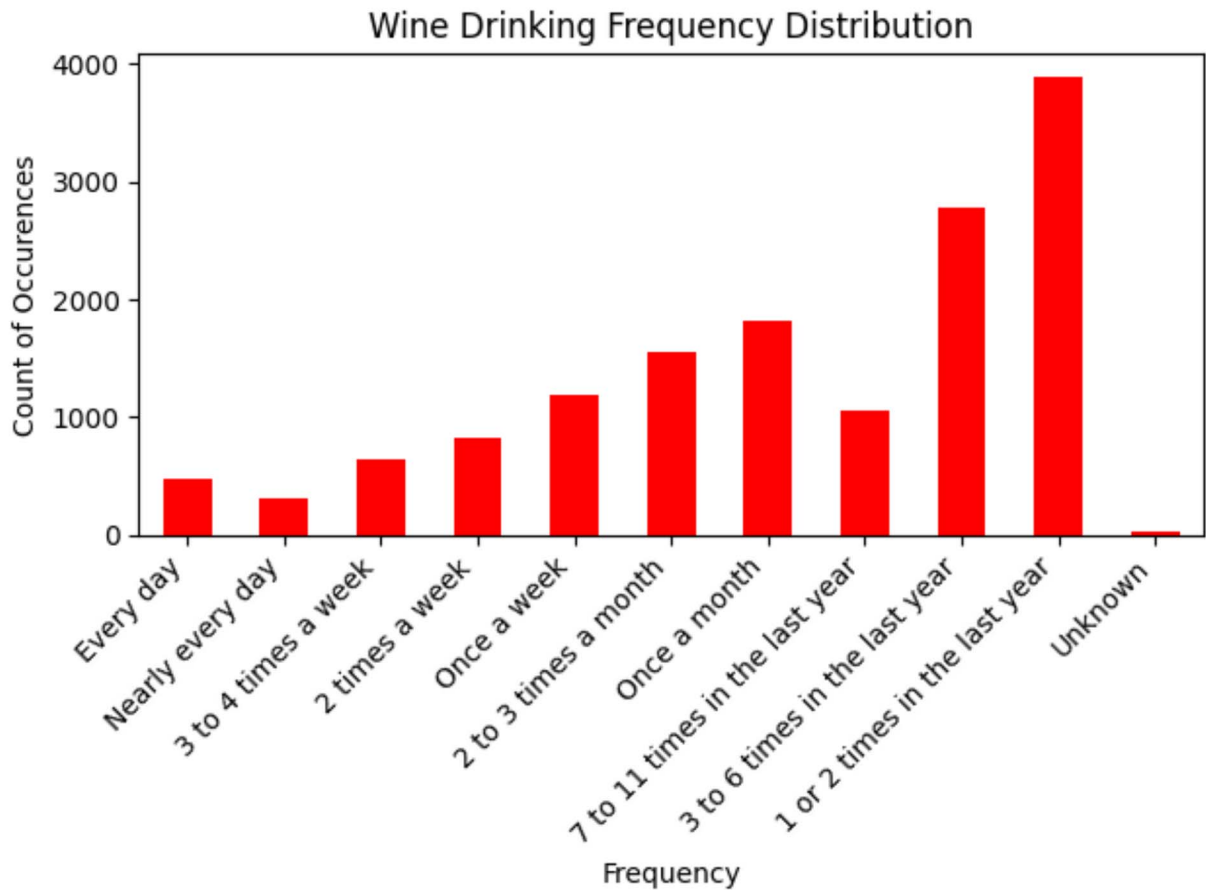


The household income is more or less equally distributed over the whole range of incomes. Nevertheless, there is a peak for incomes 40,000 dollars to 49,999 dollars, as well as for 60,000 dollars - 69,999 dollars.

```
In [4]: # Plotting the Wine Drinking Frequency Distribution
wine_freq_labels = ['Every day', 'Nearly every day',
                    '3 to 4 times a week', '2 times a week',
                    'Once a week', '2 to 3 times a month',
                    'Once a month', '7 to 11 times in the last year',
                    '3 to 6 times in the last year', '1 or 2 times in the last year',
                    'Unknown']

wine_freq_data = df_cleaned['wine_frequency'].value_counts().reindex(wine_freq_labels)

wine_freq_data.plot(kind='bar', color='red')
plt.title('Wine Drinking Frequency Distribution')
plt.xlabel('Frequency')
plt.ylabel('Count of Occurrences')
plt.xticks(rotation=45, ha='right')
plt.tight_layout() # Adjust layout to make room for x-axis labels
plt.show()
```

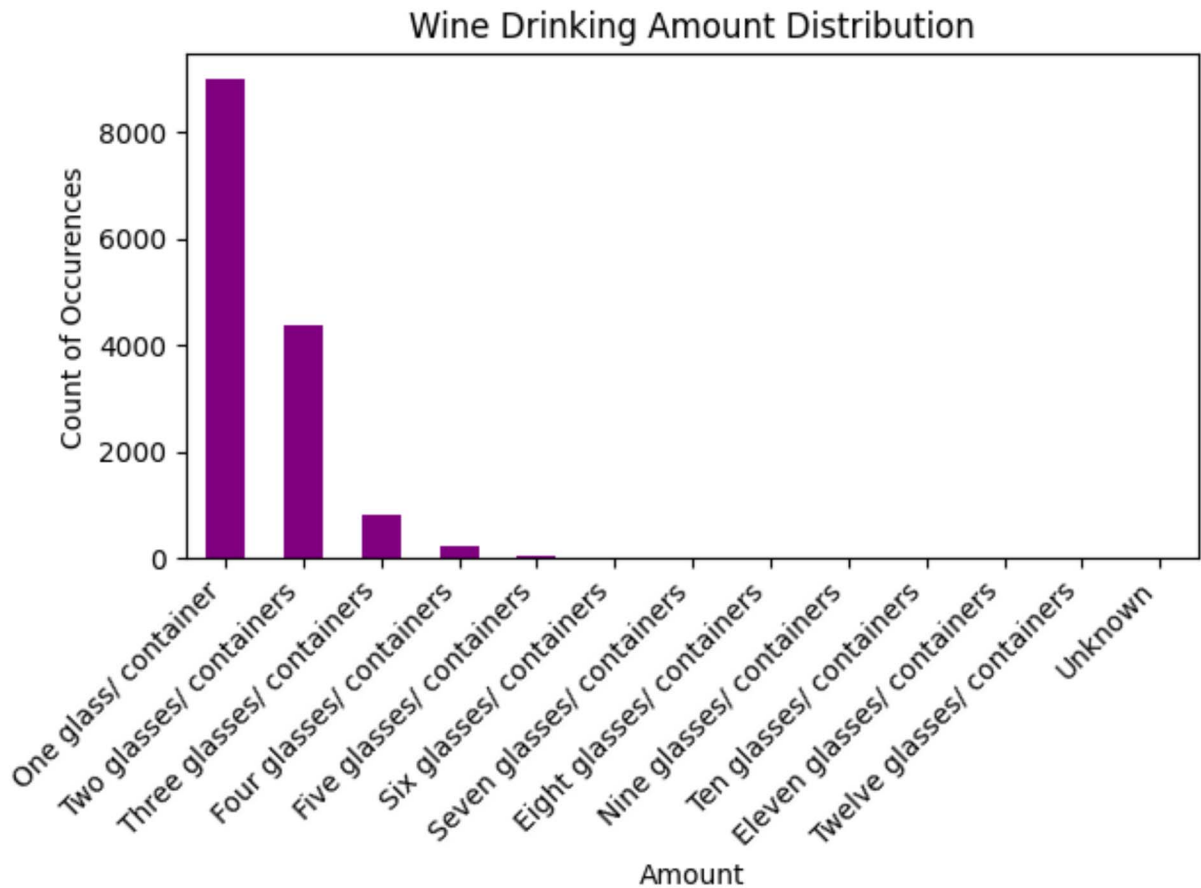


The most common category for wine drinking frequency is "1 or 2 times in the last year" followed by "3 to 6 times a year". The count of frequency is decreasing with increasing frequency.

```
In [5]: # Plotting the Wine Drinking Amount Distribution
wine_amount_labels = ['One glass/ container',
    'Two glasses/ containers',
    'Three glasses/ containers',
    'Four glasses/ containers',
    'Five glasses/ containers',
    'Six glasses/ containers',
    'Seven glasses/ containers',
    'Eight glasses/ containers',
    'Nine glasses/ containers',
    'Ten glasses/ containers',
    'Eleven glasses/ containers',
    'Twelve glasses/ containers',
    'Unknown']

wine_amount_data = df_cleaned['wine_amount'].value_counts().reindex(wine_amount_labels)

wine_amount_data.plot(kind='bar', color='purple')
plt.title('Wine Drinking Amount Distribution')
plt.xlabel('Amount')
plt.ylabel('Count of Occurences')
plt.xticks(rotation=45, ha='right')
plt.tight_layout() # Adjust Layout to make room for x-axis labels
plt.show()
```

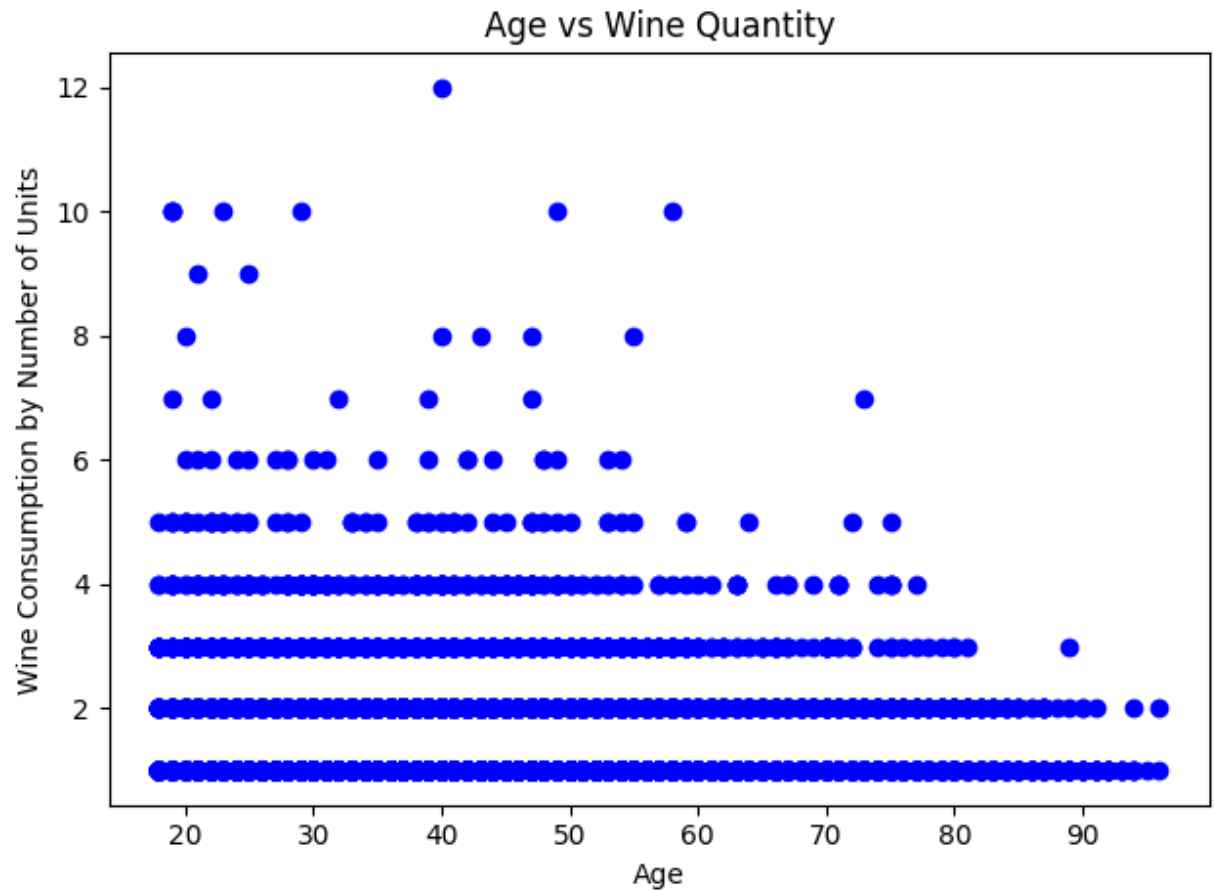


When drinking wine, the most common amount of wine drank, is 1 glass, followed by 2 glasses or 3 glasses. Higher amounts of wine drank are rare, but also present in the data.

```
In [6]: # Remove the 99 category ("Unknown") from the data, since it does not benefit the analysis
df_cleaned = df_cleaned[~(df_cleaned['noofwines'] == 99)]

# Create scatter plot
plt.scatter(df_cleaned_2['age'], df_cleaned_2['noofwines'], color='blue', marker='o')
plt.title('Age vs Wine Quantity')
plt.xlabel('Age')
plt.ylabel('Wine Consumption by Number of Units')

plt.tight_layout()
plt.show()
```



When looking at the wine consumption amount over age, it becomes clear, that the amount of wine is reducing with increasing age. The data is not randomly distributed and a pattern can be recognized. Further, some amounts, e.g., 1 and 2 glasses can be found in every age group or age. Heavy drinking in terms of the amount can only be observed for ages up to approximately 50 with 6 or more glasses per occasion.

Summary:

The data provided, seems to be a good sample for the analysis of wine consumption. Nevertheless, some age groups or income groups are more represented than others.