In [19]:

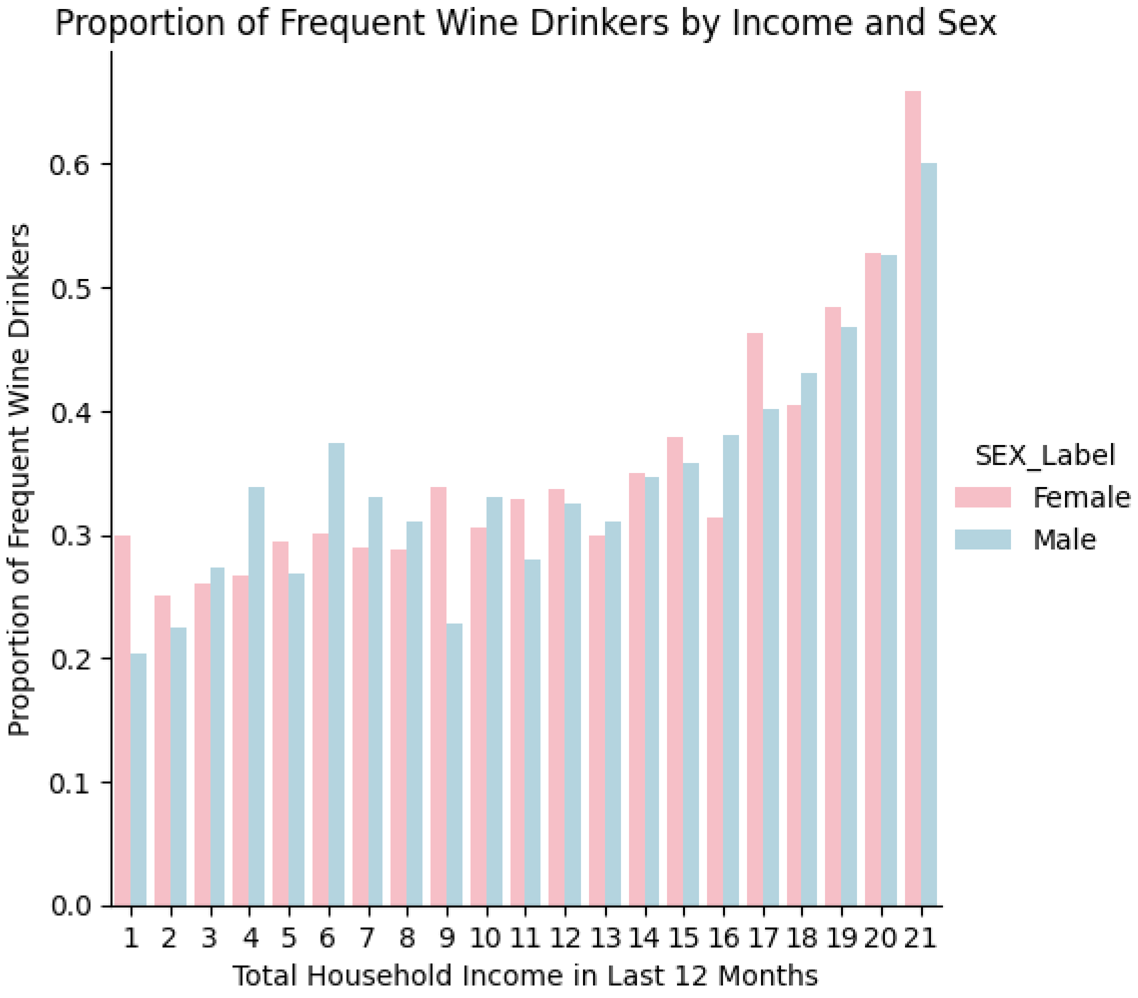
```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import seaborn as sns
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt


#######################
# Data Pre-Processing #
#######################

# Executing the same steps, as done for the linear regression models

# Bug fix for display formats to avoid run time errors
pd.set_option('display.float_format', lambda x:'%.2f'%x)

# read the data as pandas data frame
data = pd.read_csv('nesarc_pds.csv', low_memory=False)

# Create a list of the keys I would like to keep
keys_to_keep = [
    "S1Q12B",  # TOTAL HOUSEHOLD INCOME IN LAST 12 MONTHS: CAT
    "AGE",     # Age
    "SEX",     # Sex
    "S2AQ6B",  # Frequency of wine consumption in last 12 months
    "S2AQ6D"   # Number of wines drank when drinking in last 12 months
]

# Filter the data frame to keep only the specified columns
data = data[keys_to_keep]

# Setting variables to numeric
data['S1Q12B'] = pd.to_numeric(data['S1Q12B'], errors = 'coerce') # Household Income
data['S2AQ6B'] = pd.to_numeric(data['S2AQ6B'], errors = 'coerce') # Wine drinking fre
data['S2AQ6D'] = pd.to_numeric(data['S2AQ6D'], errors = 'coerce') # Wine drinking amo
data['AGE'] = pd.to_numeric(data['AGE'], errors = 'coerce') # Age
data['SEX'] = pd.to_numeric(data['SEX'], errors = 'coerce') # Sex

# Replace all '99' values by 'NaN', since they do not benefit the evaluation
data['S2AQ6B'] = data['S2AQ6B'].replace(99, np.nan)
data['S2AQ6D'] = data['S2AQ6D'].replace(99, np.nan)

# Drop rows where there are NaN values in columns 'S2AQ6B' or 'S2AQ6D'
data = data.dropna(subset = ['S2AQ6B', 'S2AQ6D'])

# Renaming of the columns for more obvious interpretation
data = data.rename(columns = {'S1Q12B': 'Household_Income', 'S2AQ6B': 'Winefrequency
```

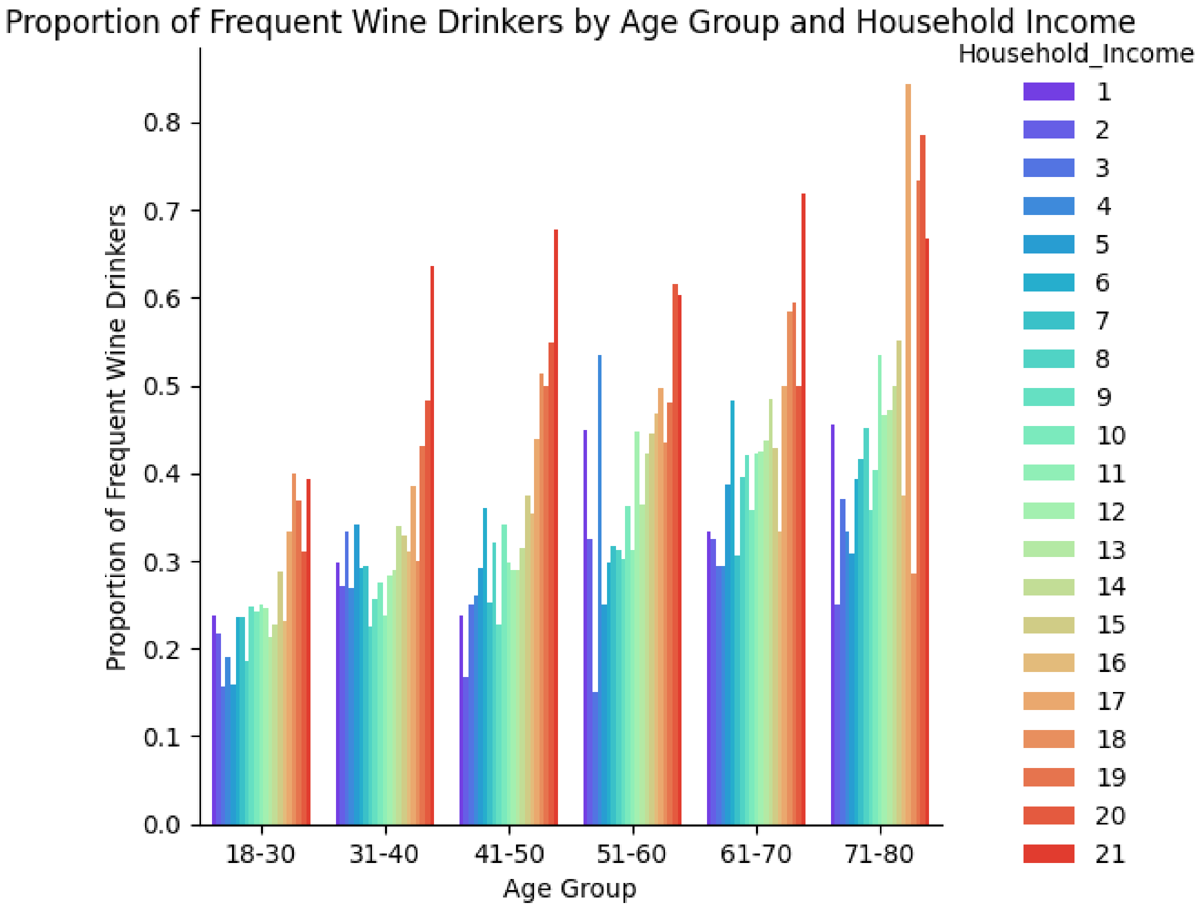
Since the drinking amount seems to be constant (based on my linear regression model results), I would like to focus now on drinking frequency. For this, I defined the border to "bad drinking habits" as drinking more often than once a month. This categorization captures a broader range of wine drinkers, including those who may drink wine occasionally but not frequently. The model will help identify factors associated with any level of wine consumption, which could be relevant for understanding general drinking habits.

In [20]:

```
# Focus on people with problematic drinking habits, which drink wine more often than
# Remember the coding of the NESARC Code Book:
# 1. Every day
# 2. Nearly every day
# 3. 3 to 4 times a week
# 4. 2 times a week
# 5. Once a week
# 6. 2 to 3 times a month
# 7. Once a month
# 8. 7 to 11 times in the last year
# 9. 3 to 6 times in the last year
# 10. 1 or 2 times in the last year
#
data['Drinks_Wine'] = (data['Winefrequency'] < 7).astype(int)

#######################
# Logistic Regression #
#######################

# Logistic Regression for Drinks_Wine ~ Household_Income + AGE + SEX
logistic_model = smf.logit('Drinks_Wine ~ Household_Income + AGE + SEX', data = data)

# Print the result
print(logistic_model.summary())

##############################################
# Plotting of data for better interpretation #
##############################################

# Listwise deletion for calculating means for regression model observations
data_cleaned = data[['Drinks_Wine', 'Household_Income', 'AGE', 'SEX']].dropna()

# Group means and standard deviations
mean_values = data_cleaned.groupby(['Household_Income', 'SEX']).mean()['Drinks_Wine']
std_values = data_cleaned.groupby(['Household_Income', 'SEX']).std()['Drinks_Wine']

# Map numeric values to labels for the 'SEX' variable
data_cleaned['SEX_Label'] = data_cleaned['SEX'].map({1: 'Male', 2: 'Female'})

# Define a custom color palette for the SEX variable
custom_palette = {'Male': (173/255, 216/255, 230/255, 1),  # Light Blue
                  'Female': (255/255, 182/255, 193/255, 1)}  # Light Red

# Bivariate bar graph for Household_Income and Drinks_Wine
sns.catplot(x = "Household_Income", y = "Drinks_Wine", hue = "SEX_Label", data = data
plt.xlabel('Total Household Income in Last 12 Months')
plt.ylabel('Proportion of Frequent Wine Drinkers')  # Updated label
plt.title('Proportion of Frequent Wine Drinkers by Income and Sex')  # Updated title
plt.show()

# Create age bins
data_cleaned['AGE_Bin'] = pd.cut(data_cleaned['AGE'], bins=[18, 30, 40, 50, 60, 70, 8

# Create a rainbow color palette based on the number of unique Household Income categ
rainbow_palette = sns.color_palette("rainbow", n_colors = data_cleaned['Household_Inc

# Bivariate bar graph for AGE_Bin and Drinks_Wine, colored by Household_Income
sns.catplot(x = "AGE_Bin", y = "Drinks_Wine", hue = "Household_Income", data = data_c
plt.xlabel('Age Group')
plt.ylabel('Proportion of Frequent Wine Drinkers')
plt.title('Proportion of Frequent Wine Drinkers by Age Group and Household Income')
plt.show()
```

```
Optimization terminated successfully.
         Current function value: 0.622744
         Iterations 5
                           Logit Regression Results                           
==============================================================================
Dep. Variable:            Drinks_Wine   No. Observations:                14522
Model:                          Logit   Df Residuals:                    14518
Method:                           MLE   Df Model:                            3
Date:                Tue, 01 Apr 2025   Pseudo R-squ.:                 0.03205
Time:                        15:48:25   Log-Likelihood:                -9043.5
converged:                       True   LL-Null:                       -9342.9
Covariance Type:            nonrobust   LLR p-value:                1.844e-129
====================================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept           -2.4427      0.099    -24.624      0.000      -2.637      -2.248
Household_Income     0.0661      0.004     17.246      0.000       0.059       0.074
AGE                  0.0200      0.001     18.262      0.000       0.018       0.022
SEX                  0.0621      0.036      1.705      0.088      -0.009       0.134
====================================================================================
```





Interpretation of results:

My variable "Drinks_Wine" is defined as "1 if drinks wine more often than once a month, 0 otherwise".

The Log-Likelihood: -9043.5, indicates that the model does fit to the data, but still could be better (less negative).

A Pseudo R-squared of 0.03205, suggests that the model explains about 3.2% of the variance in the dependent variable. This is relatively low, indicating that while the model has some predictive power but is not able to explain the complete data. This indicates, that there might be confounding for the association between your primary explanatory variable and the response variable.

The Statistical Significance is given by the LLR p-value which is 1.844e-129. This is extremely small and indicates that at least one of the predictors is statistically significant in predicting the outcome.

Coefficients Interpretation

- Household Income (Coefficient: 0.0661):

  For each one-unit increase in Household_Income, the log odds of drinking wine more often than once a month increases by 0.0661, holding all other variables constant. This indicates, that the wine drinking frequency increases with increasing income, or it is more likely to drink more frequently, if being in a upper income category.

- Age (Coefficient: 0.0200) For each additional year of age, the log odds of drinking wine more often than once a month increases by 0.0200, holding all other variables constant. This means, that more frequent drinking is associated with being in the higher ones, of the age groups.
- Sex (Coefficient: 0.0621): The coefficient for SEX indicates that being male is associated with an increase in the log odds of drinking wine more often than once a month by 0.0621, holding all other variables constant. This means that males have approximately 6.21% higher odds of drinking wine more often than once a month compared to females.

Statistical Significance:

- Household Income and Age: Both variables are statistically significant predictors of drinking wine more often than once a month, with positive coefficients indicating that higher income and older age are associated with increased likelihood of more frequent wine consumption.
- Sex: While the coefficient for SEX is positive, indicating that males are more likely to drink wine more frequently than females, it is not statistically significant at the 5% level.

Interpretation with respect to my initial hypotheses:

- H1: "Drinking of wine increases with increasing income"
    - The model supports the hypothesis, since it describes an increase of wine drinking amount with increasing income, for every age group. Also, the income variable is judged as statistically significant, which indicates that the hypothesis can be confirmed.
- H2: "The amount of consumed wine is independent of age"
    - The model does not support this hypothesis, since it shows a significant increase of wine drinking amount with increaseing age. Further, the age variable is judged as statistically significant, which indicates, that we can reject the hypothesis.