

```
In [ ]: import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.formula.api import ols

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

print(df_cleaned)
```

Recalling my two hypotheses from Module 1 Week 1:

Hypothesis 1: "Drinking of wine increases with increasing income"

Hypothesis 2: "The amount of consumed wine is independent of age"

```
In [ ]: # Performing ANOVA analyses
# ANOVA for Hypothesis 1: "Drinking of wine increases with increasing income"
model_h1 = ols('noofwines ~ C(income_category)', data=df_cleaned).fit()
anova_table_h1 = sm.stats.anova_lm(model_h1, typ=2)
print("ANOVA Result for Hypothesis 1")
print(anova_table_h1)
```

Output:

ANOVA Result for Hypothesis 1				
	sum_sq	df	F	PR(>F)
C(income_category)	573.804869	20.0	1.379763	0.119563
Residual	302339.033947	14540.0	NaN	NaN

Assessment of the output:

Since the p-value (0.1196) is greater than the common significance level of 0.05, which means that I fail to reject the null hypothesis. This means that there is not enough evidence to conclude that wine consumption significantly varies with different income categories.

```
In [ ]: # ANOVA for Hypothesis 2: "The amount of consumed wine is independent of age"
model_h2 = ols('noofwines ~ C(age_group)', data=df_cleaned).fit()
anova_table_h2 = sm.stats.anova_lm(model_h2, typ=2)
print("ANOVA Result for Hypothesis 2")
print(anova_table_h2)
```

Output:

ANOVA Result for Hypothesis 2				
	sum_sq	df	F	PR(>F)
C(age_group)	237.184215	8.0	1.425414	0.179986
Residual	302675.654601	14552.0	NaN	NaN

Assessment of the output:

Since the p-value (0.1799) is greater than the common significance level of 0.05, I fail to reject the null hypothesis. This means that there is not enough evidence to conclude that wine consumption significantly varies with different age groups.