1. Estimates of parameter values from the model you described. List the adjusted and unadjusted r^2 and give an interpretation of each. Report the results of an F-test of the model's validity at a 10% level of significance.
    a. R-squared: 0.713 → 71.3% of the variance in the sale prices can be explained by the model.
    b. Adjusted R-squared: 0.712 → the model's goodness of fit is still strong when accounting for the number of predictors.
    c. F-statistic: 902.6 with a very low p-value means the model is statistically significant at the 10% level of significance. So at least one of the independent variables is statistically significant.
2. Interpret each of the estimated coefficients. What are their signs and significance levels? Explain whether they match your expectations or how you have subsequently updated your prior?
    a. Intercept: The predicted sale price when all variables are zero is 2.731e+06. This would makes sense if there is omitted variables.
    b. bldgtype: Different building types have different effects on sale prices compared to the baseline type
        i. '2fmCon' has a coefficient of -9887.8016 means they tend to have a lower sale price compared to the baseline.
        ii. 'Duplex' houses have a coefficient of -12450, means they also tend to have a lower sale price compared to the baseline.
        iii. 'Twnhs' and 'TwnhsE' have negative coefficients means they also tend to have lower sale prices.
    c. overallqual: A higher overall quality rating is associated with a lower sale price, which is not what I would expect and could be due to omitted variables.
    d. fullbath: Each additional full bathroom is associated with an increase in sale price of 19,940.
    e. yearbuilt: For each year increase in the year the house was built, the sale price decreases by 1,422.5655.
    f. qualXyear: the difference in the marginal effects of overallqual and year build is 295.54.
    g. T.2fmCon is the only variable that is not statistically significance at a 5% significance value because the p-values < .05
3. Explain whether it is possible that your estimates are biased by omitted variables or what steps you took to investigate possible multicollinearity among the included regressors.
    a. Omitted variable bias could be present because there are more variables that affect house prices that are not included in the model.
    b. I used variance inflation factors (VIF) to assess the extent of multicollinearity among the predictors.
4. Summarize your results and use your model to predict the price of a representative good of average characteristics.
    a. The model predicts sale prices based on building type, overall quality, number of full bathrooms, year built, and the interaction between quality and year built. To predict the price of a representative house, you would substitute the values of these variables into the regression equation and solve for the sale price.

```
import pandas as pd
import numpy as np
import patsy
from patsy import dmatrices
import statsmodels.formula.api as smf
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor

pd.set_option("display.max_columns", None)
df = pd.read_stata("../data/AmesHousingGeo.dta")
df

df["qualXyear"] = df["overallqual"] * df["yearbuilt"]

y, X = patsy.dmatrices('saleprice ~
Q("overallqual")+Q("fullbath")+Q("yearbuilt")+Q("bldgtype")+Q("qualXyear")', df)
salary_model = sm.OLS(y, X)
res = salary_model.fit()
print(res.summary())

print('The parameters are:\n', res.params, '\n')
print('The confidence intervals are:\n', res.conf_int(), '\n')
print('The r-squared is:', res.rsquared)

y, X = patsy.dmatrices('saleprice ~
Q("overallqual")+Q("fullbath")+Q("yearbuilt")+Q("bldgtype")+Q("qualXyear")', df,
return_type='dataframe')
vif = pd.DataFrame()
vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif['variable'] = X.columns
vif
```