Peyton Bailey

April 27th, 2025

D600- Task 2

B. Describe the purpose of this data analysis by doing the following:

1. Propose one research question that is relevant to a real-world organizational situation captured in the provided dataset that you will answer using logistic regression in the initial model.

What key features predict whether a house is considered luxury in this housing market? I will explore key features such as price, square footage, number of bathrooms and bedrooms, renovation quality, garage presence, and fireplace.

2. Define one goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

The analysis aims to develop a predictive model that estimates the probability of a house being classified as a luxury property.

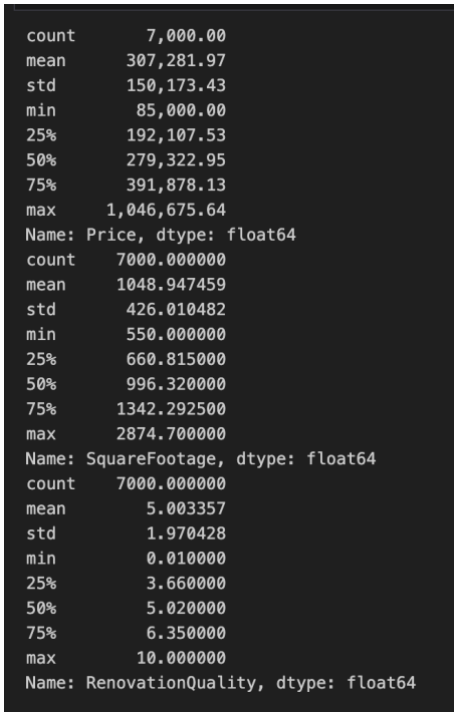
C. Summarize the data preparation process for logistic regression analysis by doing the following:

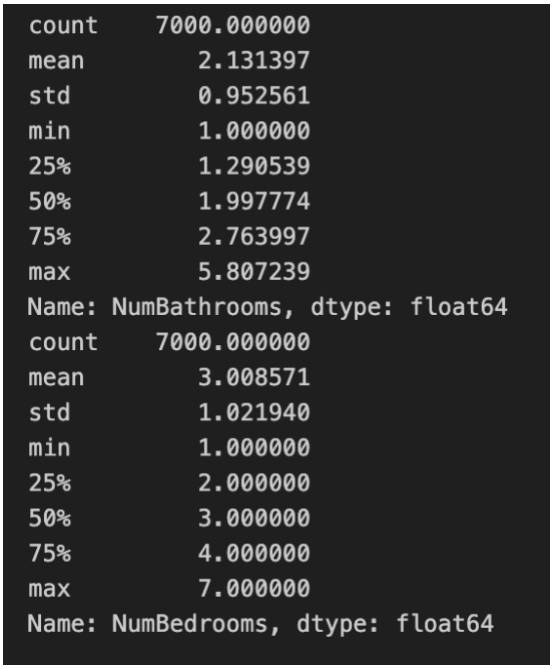
1. Identify the dependent and all independent variables that are required to answer the research question and justify your selection of variables.

The independent variables selected for my model were: Price, Square Footage, Num Bathrooms, Num Bedrooms, Renovation Quality, Garage, and Fireplace. The categorical dependent variable was Luxury. I selected these variables based on theoretical foundations in real estate valuation. Price is a fundamental indicator of property value classification. Higher-priced homes are more likely to be classified as luxury properties. Square footage is a primary determinant of real estate

value and luxury classification. Larger homes are typically more expensive and thus frequently classified as luxury. The number of bedrooms and bathrooms was included because these variables represent essential living spaces directly impacting property valuation. Properties with more bedrooms and bathrooms are associated with higher-end properties. Crime rate was selected because neighborhood safety is another critical factor in property valuation. Luxury homes are typically located in areas with lower crime rates. Renovation quality is another crucial factor because the quality of finishes and renovations significantly influences property classification. Luxury properties are characteristically known for their high-end renovations. Finally, the garage and fireplace variables were selected because these are standard features in luxury houses since they are not essential for living and often serve as an indicator of premium housing.

2. Describe the dependent variable and all independent variables from part C1 using descriptive statistics (counts, means, modes, ranges, min/max), including a screenshot of the descriptive statistics output for each of these variables.





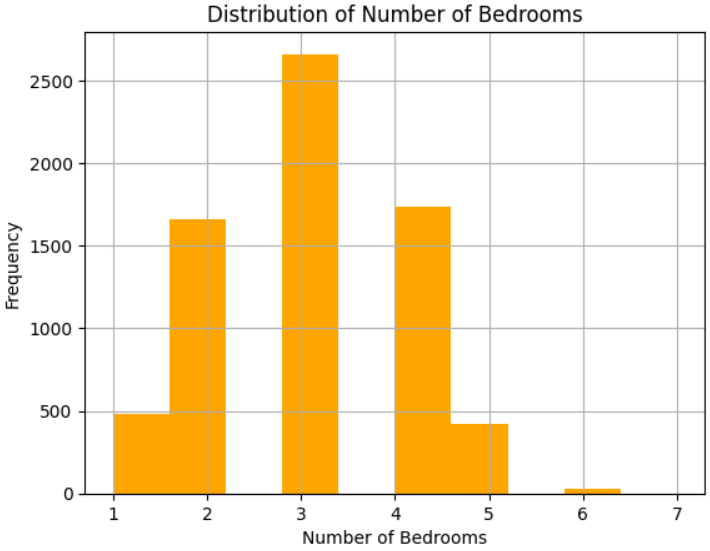
IsLuxury	Frequency	Ratios
1	3528	0.504
0	3472	0.496

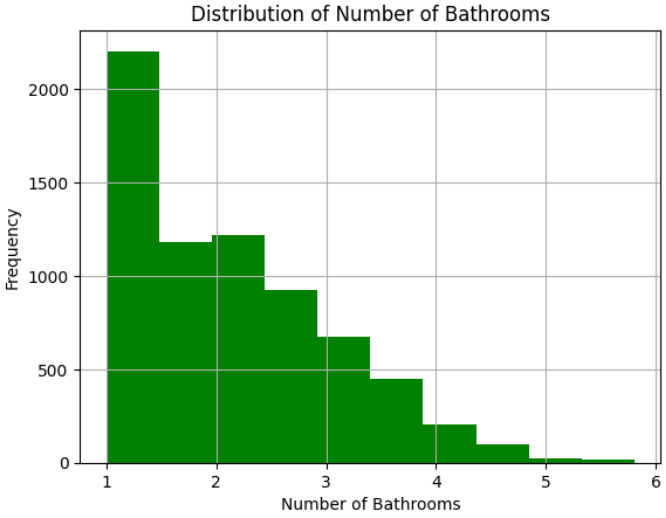
Fireplace	Frequency	Ratios
No	5172	0.738857
Yes	1828	0.261143

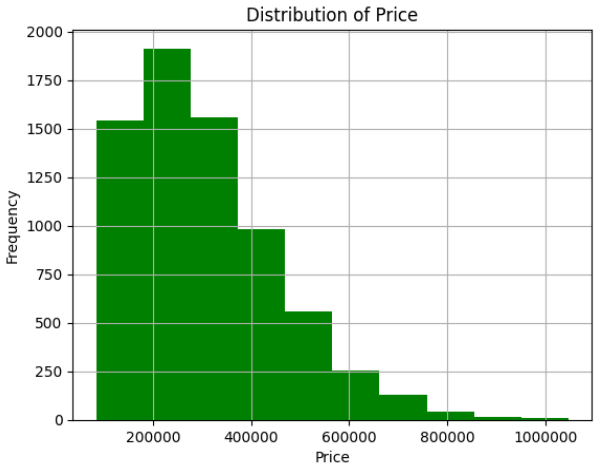
Garage	Frequency	Ratios
No	4488	0.641143
Yes	2512	0.358857

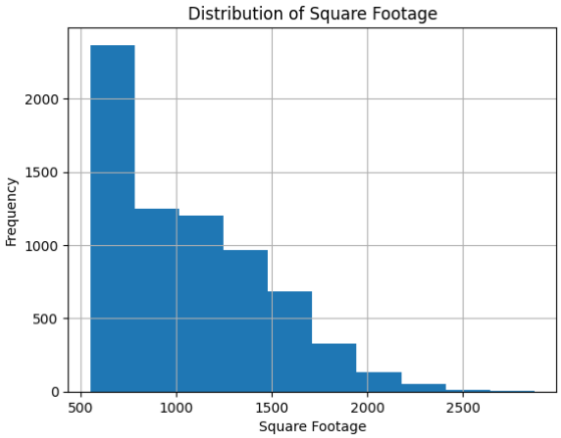
3. Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables from part C1, including the dependent variable in the bivariate visualizations.

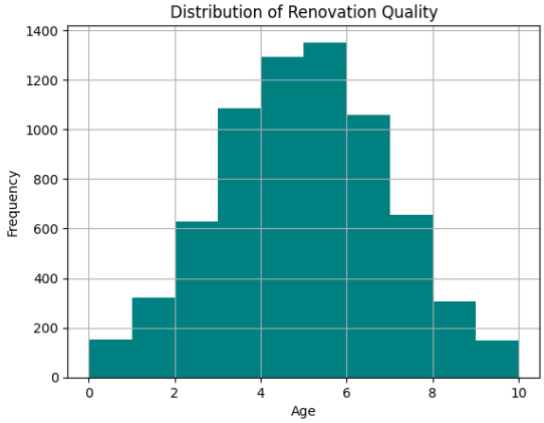
Univariate Visualizations

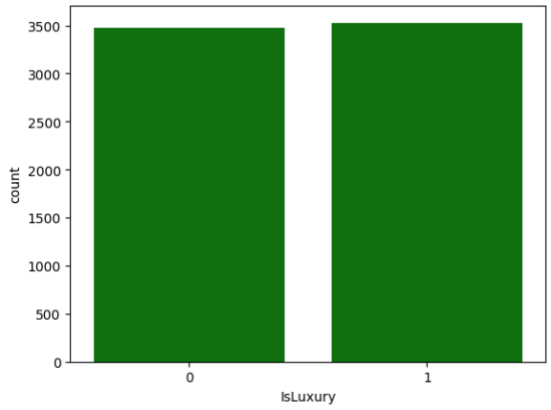


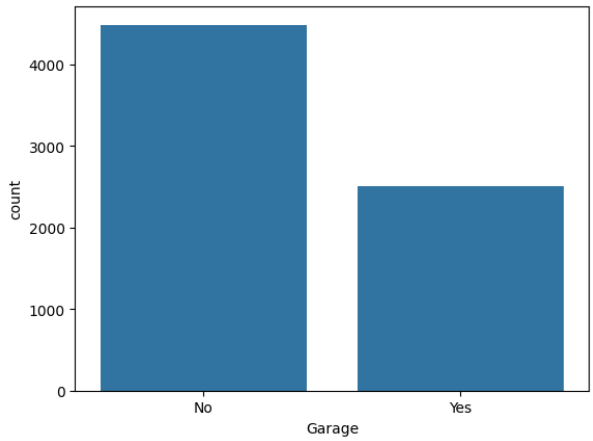


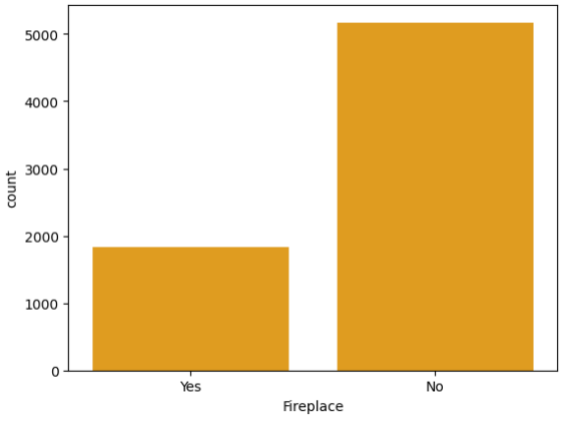

Distribution of Price
Frequency
Price
0
250
500
750
1000
1250
1500
1750
2000
200000
400000
600000
800000
1000000


Distribution of Square Footage
Frequency
Square Footage
0
500
1000
1500
2000
500
1000
1500
2000
2500

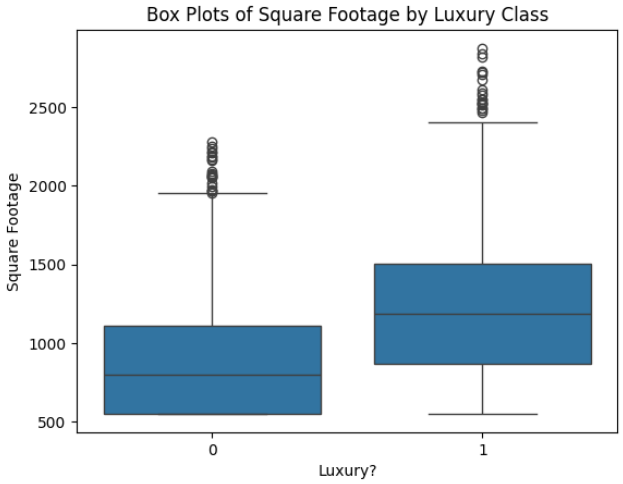

Distribution of Renovation Quality
Frequency
Age
0
200
400
600
800
1000
1200
1400
0
2
4
6
8
10

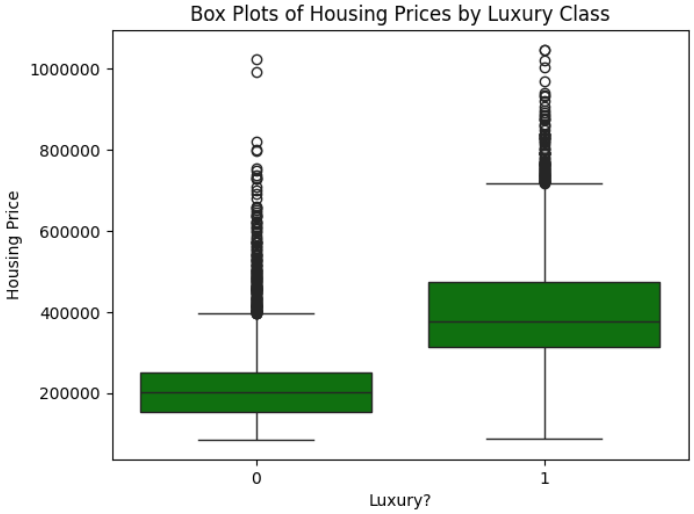

count
IsLuxury
0
500
1000
1500
2000
2500
3000
3500
0
1

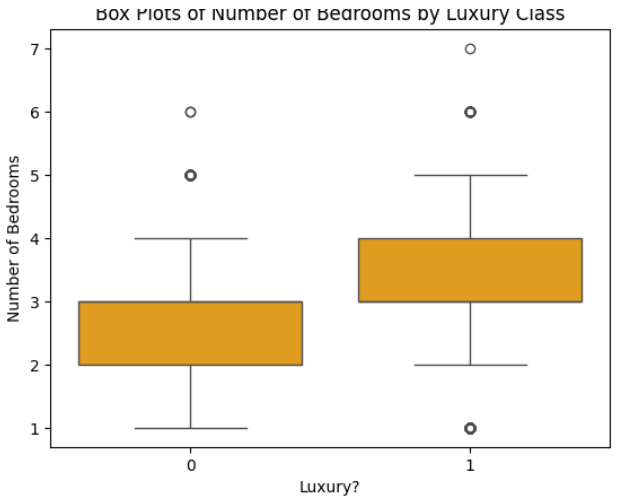

count
Garage
0
1000
2000
3000
4000
No
Yes

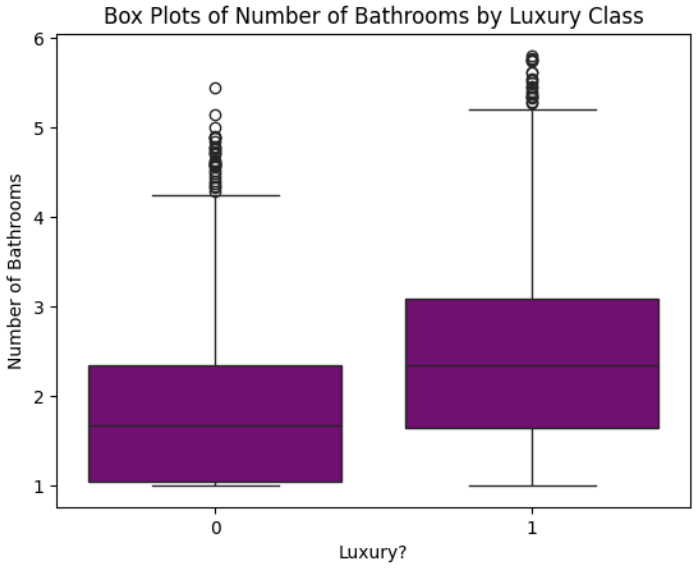

count
Fireplace
0
1000
2000
3000
4000
5000
Yes
No

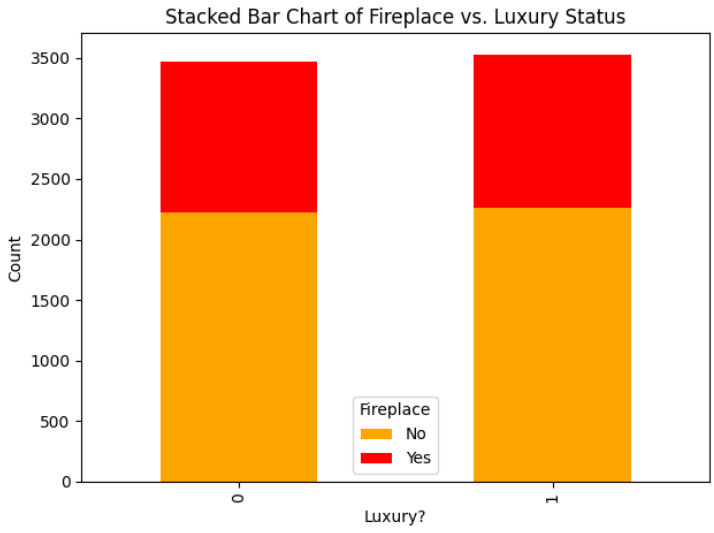
Bivariate Visualizations

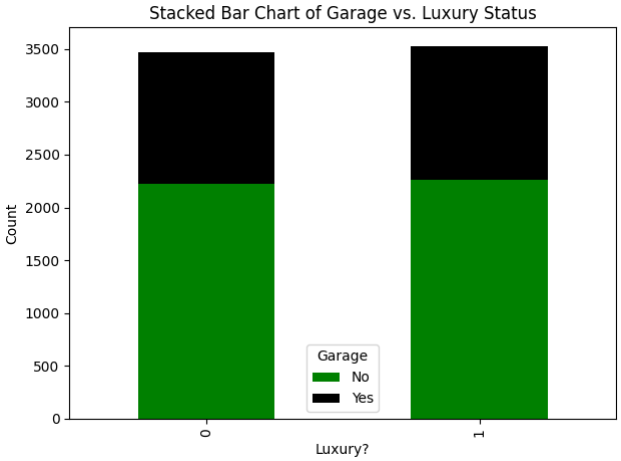












D. Perform the data analysis and report on the results by doing the following:

1. Split the data into two datasets, with a larger percentage assigned to the training dataset and a smaller percentage assigned to the test dataset. Provide the file(s).

2. Use the training dataset to create and perform a regression model using regression as a statistical method. Optimize the regression model using a process of your selection, including but not limited to, forward stepwise selection, backward stepwise elimination, and recursive selection.

Provide a screenshot of the summary of the optimized model or the following extracted model parameters:

- **AIC**
- **BIC**
- **pseudo R^2**

```
         Iterations 6
                          Logit Regression Results
==============================================================================
Dep. Variable:             IsLuxury_1   No. Observations:                 5600
Model:                          Logit   Df Residuals:                     5596
Method:                           MLE   Df Model:                            3
Date:                Sun, 27 Apr 2025   Pseudo R-squ.:                  0.2284
Time:                        15:34:17   Log-Likelihood:                -2994.9
converged:                       True   LL-Null:                       -3881.5
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          -5.6963      0.171    -33.293      0.000      -6.032      -5.361
SquareFootage   0.0020   8.34e-05     24.044      0.000       0.002       0.002
NumBathrooms    0.6611      0.036     18.478      0.000       0.591       0.731
NumBedrooms     0.7444      0.034     21.769      0.000       0.677       0.811
==============================================================================
Optimized Model Parameters:
AIC: 5997.825029578149
BIC: 6024.347117085042
Pseudo R²: 0.22840395444254058
Coefficients:
...
SquareFootage    9.604426e-128
NumBathrooms      3.101983e-76
NumBedrooms      4.590560e-105
dtype: float64
```

- **coefficient estimates**
- **p-value of each independent variable**

3. Give the confusion matrix and accuracy of the optimized model used on the training set.

```
Optimized Model Train Confusion Matrix:
 [[2117  661]
 [ 730 2092]]
Optimized Model Train Accuracy: 0.7516071428571428
```

4. Run the prediction on the test dataset using the optimized regression model from part D2 to evaluate the performance of the prediction model on the test data based on the confusion matrix and accuracy. Provide a screenshot of the results.

```
Optimized Model Test Confusion Matrix:
 [[523 171]
 [196 510]]
Optimized Model Test Accuracy: 0.7378571428571429
```

E. Summarize your data analysis by doing the following:

1. **List the packages or libraries you have chosen for Python or R and justify how each item on the list supports the analysis.**

Pandas – This package allowed me to import and manipulate the dataset within a data frame. This includes isolating each variable, calculating its descriptive statistics, and creating summary tables for the categorical variables.

Numpy – This is a fundamental library for numerical and scientific computing, which aided my analysis.

Matplotlib – This package supports the creation of visualizations. I used this package to create histograms depicting the distribution of my quantitative variables. I also used this package to visualize the relationships between categorical variables using stacked bar charts.

Seaborn – This package also supports visualizations. I used this to create box plots to depict the relationship between my quantitative explanatory variables and categorical response variable, "IsLuxury".

Statsmodels – This package directly supports regression analysis. This allowed me to add a constant to my independent variables, input the X and Y variables, and fit them into a model using Logit. This model provided a summary that included the coefficients for each independent variable, the pseudo R-squared, AIC, and BIC. Furthermore, this package also consists of the variance inflation factor (VIF), which I used to detect multicollinearity before fitting the model.

Sklearn – This package also supports regression analysis, emphasizing machine learning. Through this package, I could import a train-test split feature, allowing me to split my data into training and testing data, which allowed me to evaluate the accuracy of the prediction model. This package also allowed me to calculate the mean squared error between the predicted vs. actual values.

2. Discuss the method used to optimize the model.

This model was first checked for multicollinearity by detecting the VIF values for each independent variable. Any VIF value over 10 was considered too high. The variable with the highest VIF value was removed. The VIF values were then recalculated, and the following variable was removed. This process was completed until all VIF values were under 10. Next, I used backward stepwise elimination to eliminate variables that were not significant to the prediction of the dependent variable. This was also done one at a time using the p-value as a reference. A variable with a p-value over 0.05 was considered insignificant and thus removed from the model equation.

3. Justify the approach discussed in part E2 that was used to optimize the model.

I used backward stepwise elimination because it systematically removes variables based on their statistical significance by eliminating those that contribute the least to the model's predictive power. This reduces the model's complexity and potentially prevents overfitting. This method also makes it easy to address multicollinearity and remove variables as needed. It is also relatively straightforward and produces a decently accurate and interpretable model.

4. Summarize at least *four* assumptions of logistic regression.

The first assumption of logistic regression is that it has an appropriate outcome type. Logistic regression equations have a categorical response variable. By default, this response variable is binary, meaning it has only two possible unique outcomes. It is possible, however, for the response variable to have more than two possible outcomes. This assumption is met within my model since the response variable is "IsLuxury", which has two possible outcomes -- Yes (1) and No (0). Another assumption is the absence of multicollinearity. Multicollinearity corresponds to a situation where the data contain highly correlated independent variables. In turn, this weakens the statistical power of the model because it reduces the precision of the estimated coefficients (Leung, 2022). The variance inflation factor (VIF) measures the degree of multicollinearity within a set of independent variables.

During the optimization process of my model, any variables with a VIF of 10 or higher were removed, thus verifying this assumption. Logistic regression also assumes independence of observations, meaning they should not come from repeated or paired data. This assumption is verified via a residual series plot. This plot visualizes the deviance residuals of the model against the number of observations. The residual series plot for my optimized model was completely random, allowing me to verify this assumption. Finally, logistic regression also assumes that the dataset contains a sufficiently large sample size. A total number of observations

greater than 500 is considered a large sample size. My dataset includes 7000 observations, which satisfies this requirement to verify this assumption.

5. Provide evidence that the assumptions from part E4 were verified by providing either a code snippet or a screen shot.

Outcome Type

```
print(df['IsLuxury_1'].unique())
```

✓ 0.0s

```
[0 1]
```

Absence of Multicollinearity

```
# Provide VIF values
X_opt = X_train[variables]

# Calculate VIF for each feature
vif_data = pd.DataFrame()
vif_data['feature'] = X_opt.columns
vif_data['VIF'] = [variance_inflation_factor(X_opt.values, i) for i in range(X_opt.shape[1])]

print(vif_data)
```

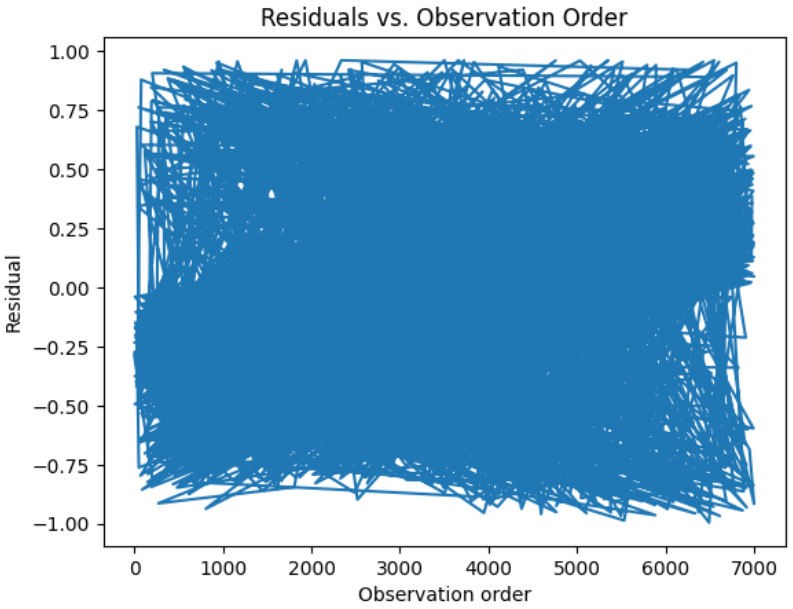
✓ 0.1s Open 'vif_data' in Data Wrangler

	feature	VIF
0	SquareFootage	5.692973
1	NumBathrooms	5.171180
2	NumBedrooms	5.978887

Independence of Observations

```
# Get residuals from statsmodels
residuals = optimized_model.resid_response

plt.plot(residuals)
plt.xlabel('Observation order')
plt.ylabel('Residual')
plt.title('Residuals vs. Observation Order')
plt.show()
```



Large Sample Size

```
print(len(df))
```

✓ 0.0s

```
7000
```

6. Provide the regression equation and discuss the coefficient estimates

```
logit(p) =  - 5.6963*const + 0.0020*SquareFootage + 0.6611*NumBathrooms + 0.7444*NumBedrooms
```

Each coefficient represents the expected change in the log-odds of a property being classified as luxury for a one-unit increase in the corresponding variable, holding all other variables constant. The SquareFootage, NumBathrooms, and NumBedrooms variables each contain a positive coefficient indicating that increases in these features are associated with higher odds of a property being luxury. The negative intercept reflects a very low baseline probability of luxury classification when all the predictors are zero. A house with no square footage, bathrooms, or

bedrooms would not be classified as luxury or even exist, for that matter, so the probability is essentially zero. Although this is not a realistic scenario in practice, it is necessary to include this for the model to be mathematically complete. A detailed explanation of the odds ratios for each variable is further discussed in section E8, along with the results and overall implications of my prediction analysis.

7. Discuss the model metrics by addressing each of the following:

- **the accuracy for the test set**

The test set achieved an accuracy of 73.8%, indicating that the model correctly classified nearly three-quarters of all properties in the unseen data.

- **the comparison of the accuracy of the training set to the accuracy of the test set**

The training set had an accuracy of 75.2%, slightly higher than the test set's accuracy of 73.8%. This suggests a well-balanced model with no signs of overfitting.

The model generalizes well to new data, which is ideal for predictive purposes.

- **the comparison of the confusion matrix for the training set to the confusion matrix of the test set**

The proportional distribution of classification results seen below is consistent between training and test sets, further confirming the model's stability and generalizability.

Training Set Confusion Matrix

- True Negatives: 2,117 (correctly identified non-luxury properties)
- True Positives: 2,092 (correctly identified luxury properties)
- False Negatives: 730 (luxury properties incorrectly classified as non-luxury)
- False Positives: 661 (non-luxury properties incorrectly classified as luxury)

Test Set Confusion Matrix

- True Negatives: 523 (correctly identified non-luxury properties)
- True Positives: 510 (correctly identified luxury properties)
- False Negatives: 196 (luxury properties incorrectly classified as non-luxury)
- False Positives: 171 (non-luxury properties incorrectly classified as luxury)

8. Discuss the results and implications of your prediction analysis.

The logistic regression model identified three significant predictors of luxury status in properties: The implications of these coefficients, when converted to odds ratios, reveal:

- **Square Footage**: For each additional square foot, the odds of a property being classified as luxury increase by a factor of 1.00 (0.2% increase per square foot)
- **Number of Bedrooms**: Each additional bedroom increases the odds of luxury classification by a factor of 2.11 (111% increase)
- **Number of Bathrooms**: Each additional bathroom increases the odds of luxury classification by a factor of 1.94 (94% increase)

These findings suggest that all three factors contribute to luxury classification. Upon first glance, one could assume that the number of bedrooms has the most substantial impact, followed closely by bathrooms, while square footage has a more modest effect on a per-unit basis. However, square footage must be interpreted in more meaningful increments since houses have hundreds or thousands of square feet. In that case,

- 100 sq ft increase: odds increase by 22% ($\exp(0.0020\times100) \approx 1.22$)
- 500 sq ft increase: odds increase by 171% ($\exp(0.0020\times500) \approx 2.71$)
- 1,000 sq ft increase: odds increase by 639% ($\exp(0.0020\times1000) \approx 7.39$)

Therefore, the house size (square footage) has the most significant overall effect on luxury classification. At the same time, the number of bedrooms and bathrooms also makes a meaningful impact on a per-unit basis.

9. Recommend a course of action for the real-world organizational situation from part B1 based on your results and implications discussed in part E8.

Real estate developers specializing in luxury-class homes should focus on building large square-footage houses with a substantial number of bedrooms and bathrooms. Adding each bedroom or bathroom doubles the odds of a luxury classification. Furthermore, each addition of 500 square feet nearly triples these odds. Ultimately, focusing on these features during the development process will allow builders and investors to make a return on their investment by creating luxury-class homes that yield high profits.

REFERENCES

Leung, K. (2022, October 4). *Assumptions of logistic regression, clearly explained*. Towards Data Science. https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290