Western Governors University

D214 – Data Analytics Graduate Capstone – Task 3

By Krista Moik

Table of Contents

Part I. Executive Summary and Implications

A. Executive Summary and Implications

Research Question and Hypotheses

Utilizing a dataset obtained from Kaggle composed of house sale data from Realtor.com, the research question for this study: **Can a multiple linear regression model be constructed based on the research dataset to determine which variables are most significant to the sale price of homes in Florida?**

The hypotheses for this research question are:

> **Null hypothesis** - H0: A predictive MLR model cannot determine which variables are most significant to the sale price of homes in Florida with an r-squared value of 0.5 or less.
>
> **Alternate Hypothesis** - H1: A predictive MLR model can determine which variables are most significant to the sale price of homes in Florida with an r-squared value greater than 0.5.

P-values will be used to ensure only statistically significant variables are included in the model. The fit of the model will be based on the r-squared value. If the r-squared value is greater than 0.5, that indicates the model explains more than 50% of the variance that affects the sale price of homes in Florida, and the null hypothesis would be rejected. If the r-squared value of the model is less than or equal to 0.5, that indicates that model is only explaining up to 50% of the variance that affects the sale price of homes in Florida, and the null hypothesis would be accepted.

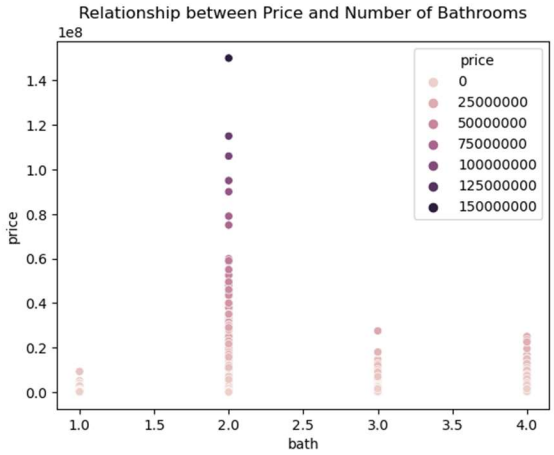
The pandemic caused many changes in the United States and the world beyond. One result of the pandemic is that many workers became remote employees and began migrating to other areas of the country they would not otherwise have had the opportunity to work from. Florida is among the states where more people are moving to and less are leaving with 2022 Census data indicating that 249,064 people moved to Florida (Davis, 2023). Due to the influx of people, the real estate landscape is heavily affected by increased demand. It is important for realtors, investors, developers, and businesses with employees moving to Florida to gain a better understanding of what factors are affecting house sale prices.

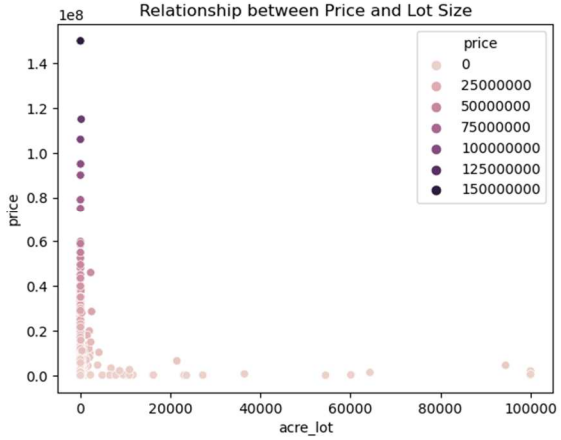
Data Analysis Process

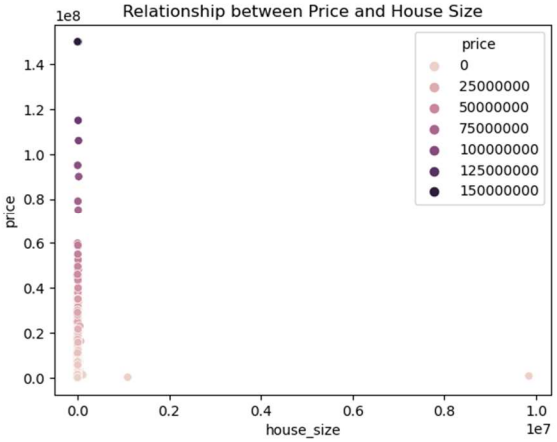
Data was first obtained from a public dataset available on Kaggle that was composed of data that had been ethically scraped from Realtor.com. The dataset consists of 10 columns that provide the sale status, sale price, number of bedrooms, number of bathrooms, lot size in acres, city, state, zip code, house size in square feet, and previously sold date. I dropped the columns for status, state, and previously sold date from my analysis.

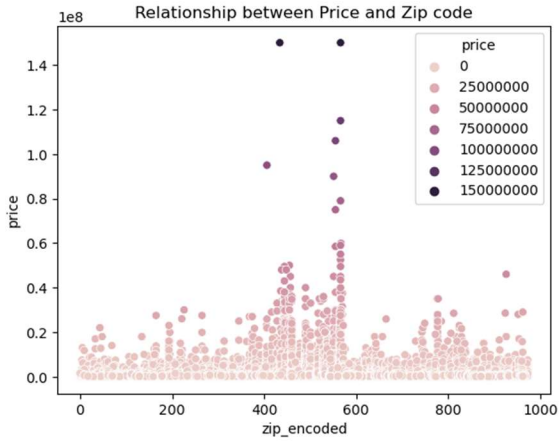
Duplicates were removed and missing values and outliers were imputed or retained as appropriate. The two categorical variables, city and zip_code, were re-expressed via LabelEncoder to assign a number to each city and zip code to include in our model. The original columns were dropped. Bivariate

visualizations were created to view relationships between each independent variable and the dependent variable of price.









The variables were fitted to the Ordinary Least Squared multiple linear regression model, resulting in an initial model:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.051
Model:                            OLS   Adj. R-squared:                  0.051
Method:                 Least Squares   F-statistic:                     1335.
Date:                Tue, 02 Apr 2024   Prob (F-statistic):               0.00
Time:                        10:14:18   Log-Likelihood:            -2.3757e+06
No. Observations:              150220   AIC:                         4.751e+06
Df Residuals:                  150213   BIC:                         4.752e+06
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -7.101e+05   2.25e+04    -31.556      0.000   -7.54e+05   -6.66e+05
bed         4.33e+05   5523.788     78.396      0.000    4.22e+05    4.44e+05
bath        2.15e+04   7885.539      2.726      0.006    6044.150     3.7e+04
acre_lot      12.5564      5.514      2.277      0.023       1.750      23.363
house_size     1.0653      0.180      5.907      0.000       0.712       1.419
city_encoded  -0.5736     22.723     -0.025      0.980     -45.110      43.963
zip_encoded   46.4232     16.605      2.796      0.005      13.877      78.969
==============================================================================
Omnibus:                   354025.305   Durbin-Watson:                   1.610
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       8818916650.502
Skew:                          23.380   Prob(JB):                         0.00
Kurtosis:                    1189.075   Cond. No.                     1.27e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.27e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The initial model showed that not all of the variables were statistically significant, and that there was also a high probability of multicollinearity. Variance Inflation Factor was used and determined there was multicollinearity between variables bed and bath, with values greater than 10. Bath, as the higher value was removed, and all remaining values had VIF values less than 10 and were kept in the model.

The city_encoded variable had a p-value greater than 0.05, indicating it was not statistically significant so, using backward stepwise elimination, it was removed from the OLS model. The final OLS model consisted of the variables bed, bath, acre_lot, house_size, and zip_encoded:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.051
Model:                            OLS   Adj. R-squared:                  0.051
Method:                 Least Squares   F-statistic:                     2000.
Date:                Tue, 02 Apr 2024   Prob (F-statistic):               0.00
Time:                        10:14:19   Log-Likelihood:            -2.3757e+06
No. Observations:              150220   AIC:                         4.751e+06
Df Residuals:                  150215   BIC:                         4.751e+06
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -6.828e+05   1.82e+04    -37.429      0.000   -7.19e+05   -6.47e+05
bed         4.398e+05   4938.185     89.058      0.000     4.3e+05    4.49e+05
acre_lot       12.5370      5.514      2.274      0.023       1.730      23.344
house_size      1.0671      0.180      5.917      0.000       0.714       1.421
zip_encoded    46.0256     16.604      2.772      0.006      13.481      78.570
==============================================================================
Omnibus:                   353916.959   Durbin-Watson:                   1.609
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       8808557581.460
Skew:                          23.362   Prob(JB):                         0.00
Kurtosis:                    1188.378   Cond. No.                     1.04e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.04e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In addition to the p-values and r-squared values, the F-statistic, AIC score, Durbin-Watson score, residual and Q-Q Plots, and MSE and RMSE were used to evaluate the fit of the model.

Findings

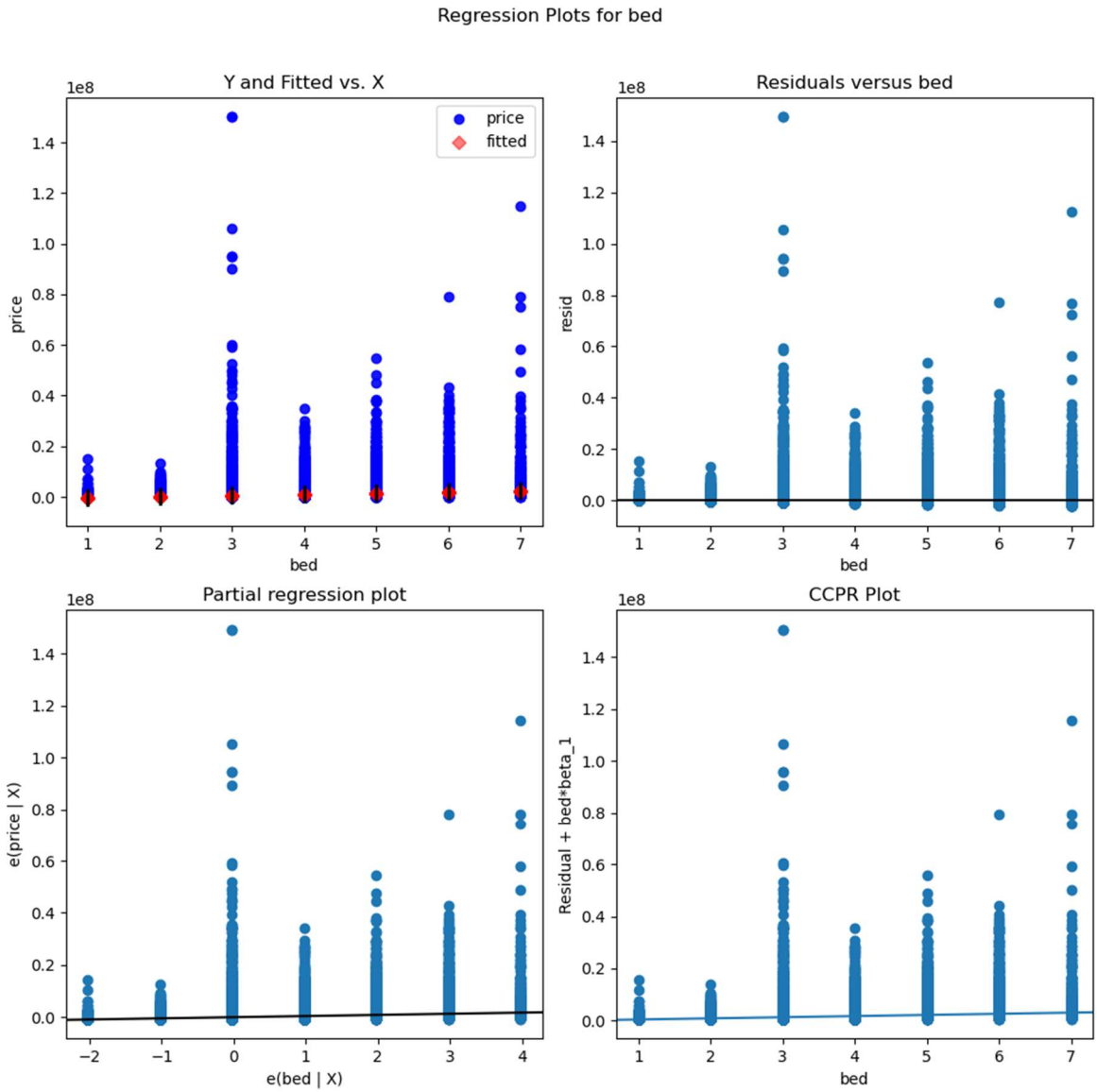
Based on the final OLS model, the final equation is: **price = -6.828e+05 + 4.398e+05 * bed + 12.5370 * acre_lot + 1.0671 * house_size + 46.0256 * zip_encoded**

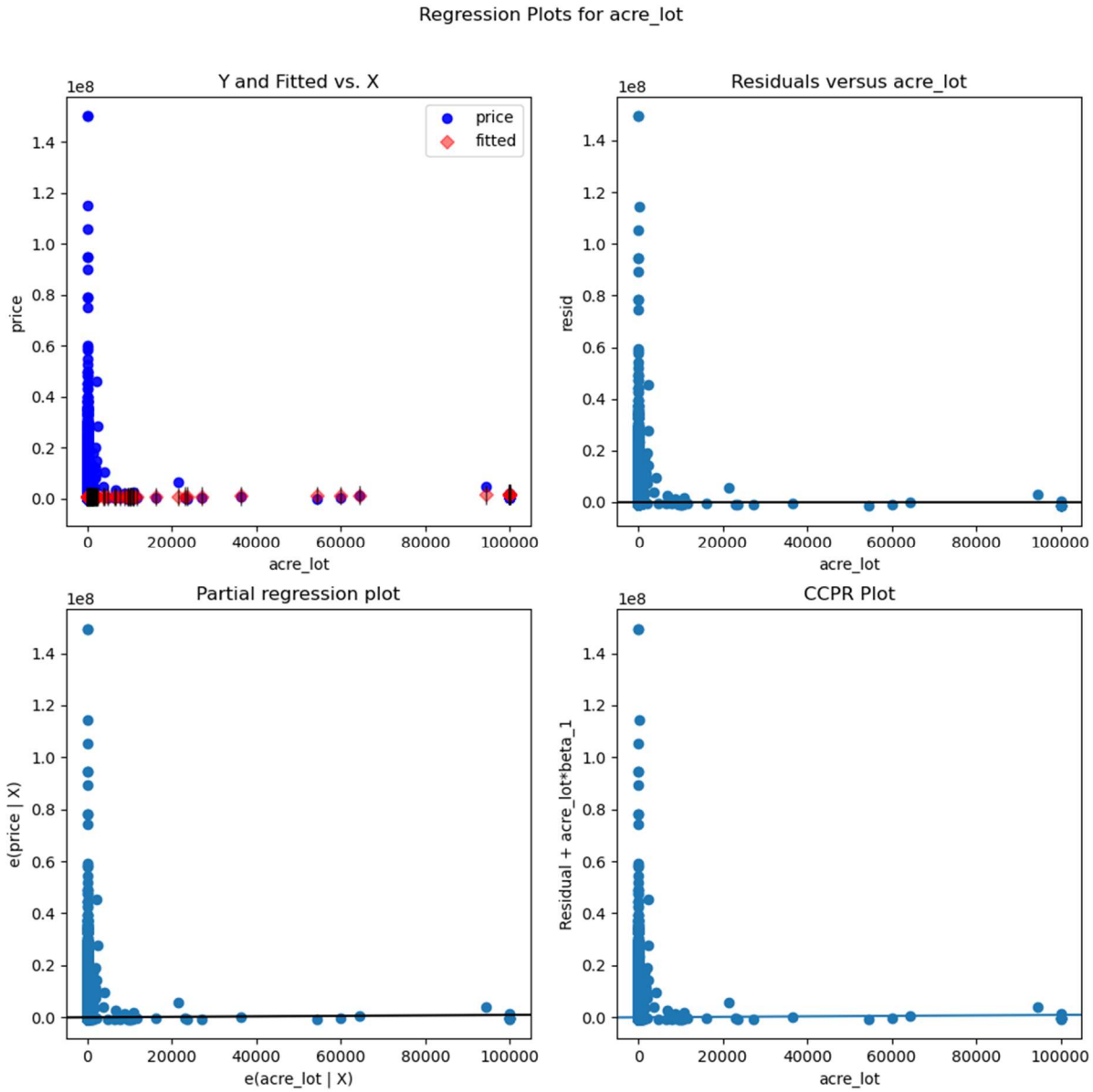
As the y-intercept, -6.828e+05 is the predicted value of the sale price of houses when all other variables are equal to 0. The coefficient of 4.398e+05 for beds means, assuming all else remains constant, an increase of 1 bedroom increases the sale price by 4.398e+05. Assuming all else remains constant, an increase of 1 unit for acre_lot results in an increase in price by 12.5370. Assuming all else remains constant, an increase of 1 unit for house_size results in an increase in price by 1.0671. Assuming all else remains constant, an increase in 1 unit for zip_encoded results in an increase in price by 46.0256.

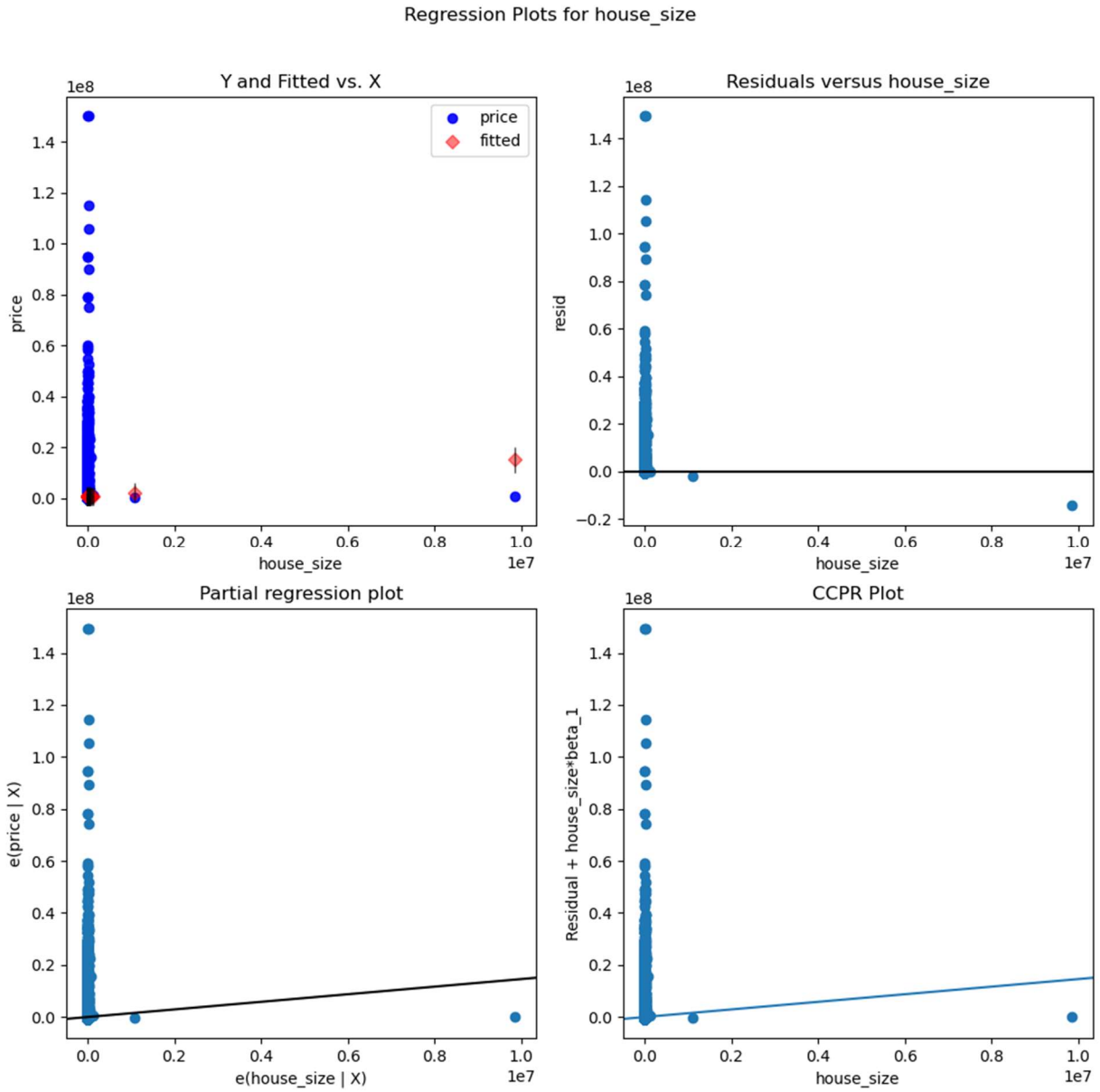
Based on this equation, the number of bedrooms is the biggest factor for the sale price of houses in Florida compared to the other independent variables. The remaining variables are still statistically significant as they all have a p-value less than 0.05.

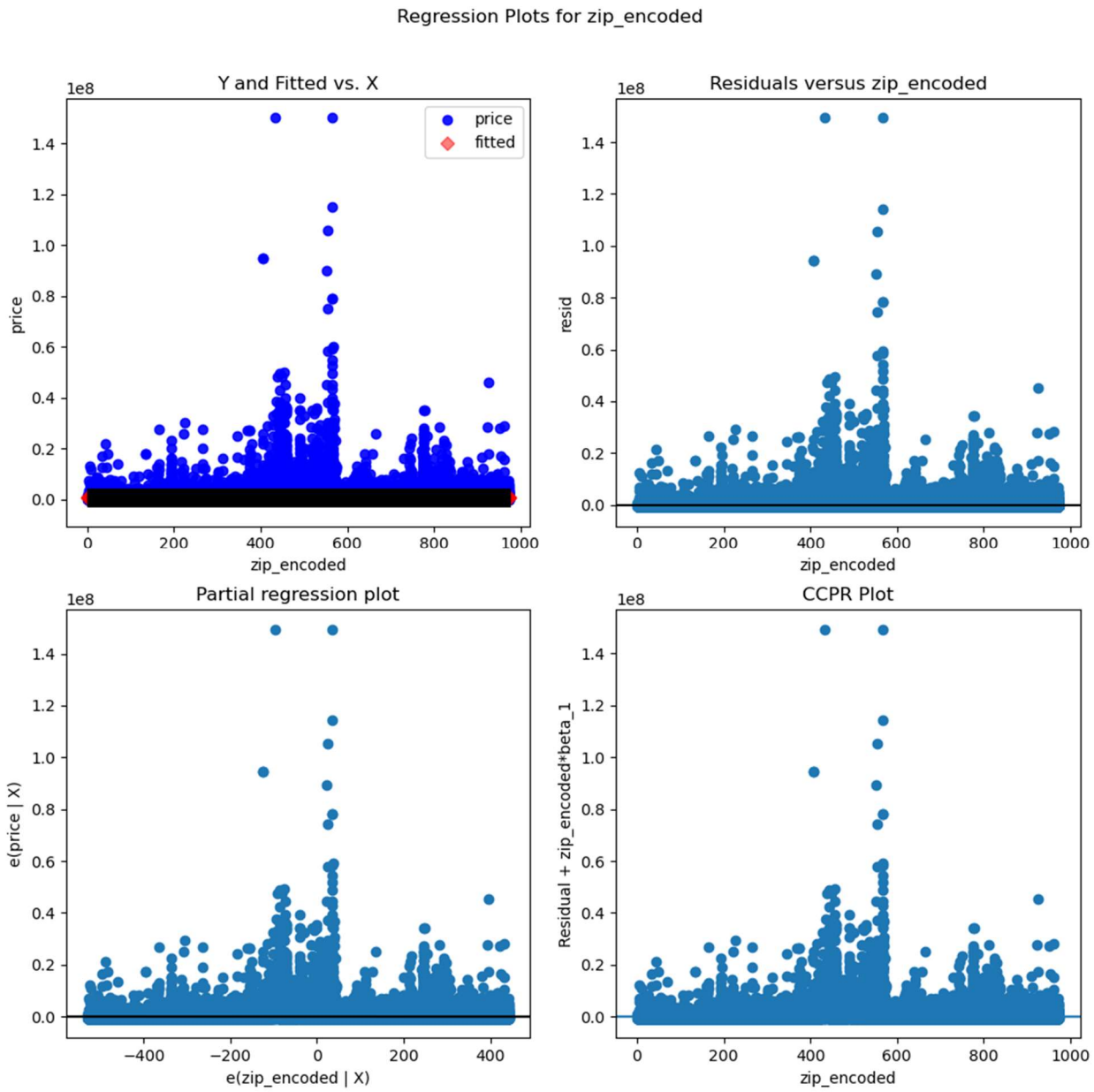
The r-squared and adjusted r-squared values of the final model are 0.051, indicating this model only explains about 5% of the variance in home sale prices in Florida. The F-statistic of 2000 indicates that this model is significant and at least one independent variable has a significant impact on the dependent variable (MathWorks, n.d.). The AIC is 4.751e+06, which is quite large, and indicates that the model may not be the best fit for the data and could be improved. The Durbin-Watson value of 1.6 indicates there is most likely not severe autocorrelation but that there does appear to be some correlation between the errors, which violates the assumption that they are independent (Medium, 2021). Thus, while my variables were determined to be statistically significant, based on the r-squared value, we must accept the null hypothesis as our model did not have an r-squared value greater than 0.5.

Residual plots for each independent variable in the final model were created. Several of the variables were unbalanced with most of the points being vertically spread around 0 on the x-axis. This confirms the model is generally accurate but indicates there could be heteroscedasticity (Qualtrics, n.d.).

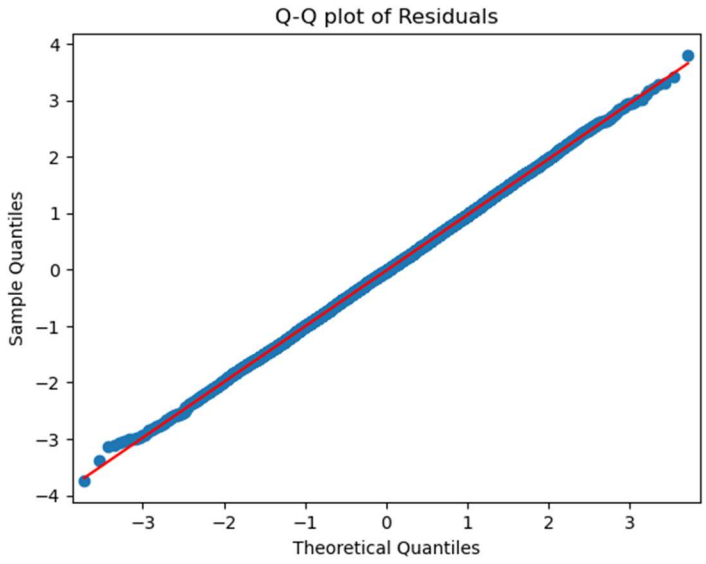



Regression Plots for acre_lot
Y and Fitted vs. X
price
fitted
1e8
price
acre_lot
Residuals versus acre_lot
resid
acre_lot
Partial regression plot
e(price | X)
e(acre_lot | X)
CCPR Plot
Residual + acre_lot*beta_1
acre_lot


Regression Plots for house_size
Y and Fitted vs. X
price
fitted
price
house_size
1e8
1e7
Residuals versus house_size
resid
house_size
Partial regression plot
e(price | X)
e(house_size | X)
CCPR Plot
Residual + house_size*beta_1
house_size



Additionally, a Q-Q Plot of the residuals was created and confirms a normal distribution, which is in line with the assumption that there is a linear relationship between the dependent and independent variables, and also indicates the model is capturing the relationships between those variables well (Frost, n.d.).



My final metric of evaluation for the model was calculating the mean squared error and root mean squared error. Both values were similar with the MSE being 0.97 and the RMSE being 0.98. This indicates the model's predictions were off by only about 1 unit per average, which is fairly accurate.

```
#mean squared error and root mean squared error of model
mse=np.mean(residuals**2)
rmse=np.sqrt(mse)
print('Mean Squared Error (MSE):', mse)
print('Root Mean Squared Error (RMSE):', rmse)
```

```
Mean Squared Error (MSE): 0.9756077773866701
Root Mean Squared Error (RMSE): 0.9877285950030353
```

Ultimately, while the most statistically significant variables were identified with their coefficients, and the model has a low rate of errors, the model itself does not explain even half of the variance that affects the sale price of houses. It is not surprising that the variables did not explain all or even most of the variance affecting sale price as there are so many other factors not accounted for in this dataset such as recent upgrades to the home, the quality of nearby schools, amenities, environmental risks, and cost of living, all of which can vary greatly in just a small area.

Limitations

The main limitation of this study and the dataset is that the dataset only provides sale data for the last two years. Additionally, more data points would be helpful in creating a better model that explained more of the variance of house sale prices. It would also be beneficial if the actual sale date was provided to determine seasonal trends, possibly via a separate Time Series study.

Other limitations of this study involve the assumptions of the model. In choosing a multiple linear regression model and ordinary least squares, several assumptions must be adhered to. First, it is assumed that the relationship between the dependent and independent variables is linear, and that the

independent variables are not strongly correlated with each other. The model also assumes that the errors are independent of each other, show homoscedasticity, and follow a normal distribution (Sampaio, 2023).

Proposed Actions

My first proposed action is to obtain additional information regarding the sold houses that would potentially affect the sale price and account for more variance within sale prices. Additionally, the final model still indicates there is possible high multicollinearity, so another variable selection tool should be explored such as PCA or ridge regression to attempt to reduce any multicollinearity that is still present. Finally, it may be beneficial to explore other model options in general such as Random Forests, which may handle the nulls and outliers better, and is less prone to overfitting while also being able to capture more complex relationships that may exist between the variables and the sale price of homes.

Benefits

The benefits of this model and understanding which factors most affect the sale price in homes, this information can be used for several different purposes:

- Real estate companies would be able to better serve their home sellers by competitively pricing the home and could better serve their home buyers by knowing the fair price of the house.
- Real estate developers and home renovation specialists would understand what increased or decreased house prices in the specified area and could tailor their work to increase profits.
- Employers can use this information to ensure that employee salaries were commensurate with the cost of living and housing in the area to attract and retain top talent for their companies.

Part II. Presentation of Findings

B. Presentation: Organization and Professionalism

B1. Presentation: Content

Please see the following link for my Panopto video recording:
https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=2fbf7aff-77e4-4c57-a799-b1510158c907

C. Sources

Davis, Elliott, Jr. (2023, December 5). These States are Bringing in More Residents than They're Losing. Retrieved April 1, 2024, from https://www.usnews.com/news/best-states/articles/these-states-are-bringing-in-more-residents-than-theyre-losing

F-statistic and t-Statistic. (n.d.). Retrieved April 1, 2024, from https://www.mathworks.com/help/stats/f-statistic-and-t-statistic.html

Frost, Jim. (n.d.). QQ Plot: Uses, Benefits & Interpreting. Retrieved April 2, 2024, from https://statisticsbyjim.com/graphs/qq-plot/

Interpreting Residual Plots to Improve Your Regression. (n.d.). Retrieved April 4, 2024, from https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

Sampaio, Vitor. (2023, June 2). Understanding Ordinary Least Squares (OLS): The Foundation of Linear Regression. Retrieved April 1, 2024, from https://medium.com/@VitorCSampaio/understanding-ordinary-least-squares-ols-the-foundation-of-linear-regression-1d79bfc3ca35

Understanding Durbin-Watson Test. (2021, August 4). Retrieved April 2, 2024, from https://medium.com/@analyttica/durbin-watson-test-fde429f79203#:~:text=The%20Durbin%20Watson%20(DW)%20statistic%20is%20used%20as%20a%20test,in%20reality%20they%20are%20not.

D. Professional Communication