Research Question

Home prices in the United States have risen substantially over the past two years. According to Reuters, the average home price has been up 17% over the last year and is expected to rise another 10.3%. (Kishan & Ganguly, 2022) The causes of this dramatic rise in prices include low-interest rates, low inventory of homes for sale, a soaring stock market, and covid. People have more cash on hand because of the lockdown. These influences have led to multiple offers above the asking price for all properties on the market. (Orton, 2022)

Knowing which features of the home are most statistically significant and influence the price would benefit investors and individuals looking to buy a home. Investors will want to maximize their return, and individuals will want to get the features they want at the best price. This study will use multiple linear regression to explore the home's features and find if these features affect the price.

The research question for this study is: Does ward affect the price of a home in Washington D.C.? What other features affect the price? The null hypothesis states that Ward and other features do not statistically significantly affect the price of a home when the alpha level is 0.05. The alternative hypothesis is that Ward and other features statistically significantly affect the home price when the alpha level is 0.05.

Data Collection

The data set for this study was downloaded from D.C. Residential Properties on Kaggle. (Chrisc, 2018) It contains 159,000 records of homes in Washington, D.C. The data set is publicly available with an Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license. The data set is comprised of 49 columns, 22 of which were used in this study. The columns used were:

|  |  |  |
| --- | --- | --- |
| Column Name | Type | Description of Data |
| BATHRM | Numeric | Number of Bathrooms |
| HF\_BATHRM | Numeric | Number of Half Bathrooms |
| HEAT | Categorical | Heating Type |
| AC | Boolean | Air Conditioning Y/N |
| NUM\_UNITS | Numeric | Number of Units |
| ROOMS | Numeric | Number of Rooms |
| BEDRM | Numeric | Number of Bedrooms |
| AYB | Numeric | When main portion of house was built |
| EYB | Numeric | Year of last improvement |
| STORIES | Numeric | Number of Stories |
| GBA | Numeric | Gross building area in square feet |
| STYLE | Categorical | Style Description |
| STRUCT | Categorical | Structure Description |
| GRADE | Categorical | Grade Description |
| CNDTN | Categorical | Condition Description |
| EXTWALL | Categorical | Exterior Wall Description |
| ROOF | Categorical | Roof Type Description |
| INTWALL | Categorical | Interior Wall Description |
| KITCHENS | Categorical | Number of Kitchens |
| FIREPLACES | Categorical | Number of Fireplaces |
| LANDAREA | Categorical | Land Area of Property in square feet |
| WARD | Categorical | City Ward |

One advantage of this data-gathering methodology is its accessibility. The data set is available for free online at Kaggle.com. It has a high useability rating from Kaggle and includes many different features for homes.

A disadvantage of this data set was that many columns have a large percentage of null values. The target variable, Price, consists of 39% nulls.

Data Extraction and Preparation

Python was used to extract and prepare for analysis. Python is an easy-to-use, open-source language that offers an extensive library of external packages to expand Python to fit the user's needs. Pandas and NumPy were used to import, clean, and transform the data set. Seaborn and Matplotlib were used to create visualizations of the data. Python does have some disadvantages. Because Python is an interpreted and dynamically typed language, it can execute code slower than other languages. Python also uses more memory. (Python Advantages and Disadvantages - Step in the right direction, n.d.)

The data was loaded into a Jupyter notebook for cleaning and analysis. A new DataFrame was created with the columns being used in the study. The following steps were used to prepare the data:

* Since PRICE is the target variable, all rows that do not have a price are removed.

Graphical user interface

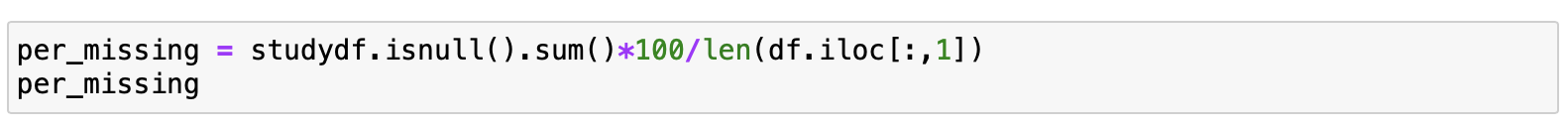
Description automatically generated

* The word ward was removed from all entries in the WARD column, and the data type was changed to integer. (Code From: https://stackoverflow.com/questions/51778480/remove-certain-string-from-entire-column-in-pandas-dataframe)

Text

Description automatically generated

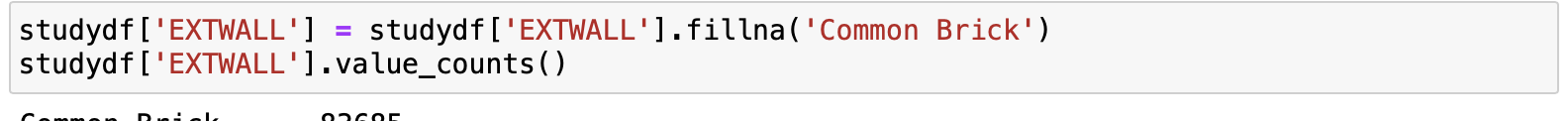
* The percent of nulls for each column was calculated. There were 11 columns with 25% of the data missing. The null values need to be filled for the columns to be kept. There are many ways to handle this. I choose a combination of imputing the mode or median and creating a 'Missing' classification.



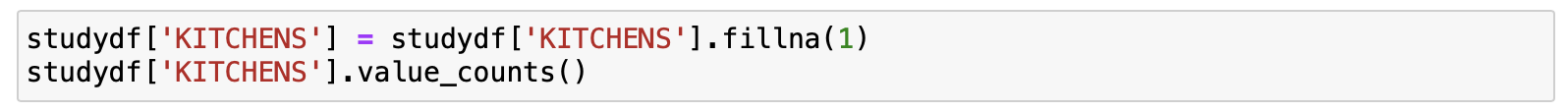
* The mode was used to replace the missing data for NUM\_UNITS, STORIES, EXTWALL, INTWALL, and KITCHENS. The mode was used because, in each case, the mode occurred much more frequently than any other. For example, NUM\_UNITS had 49339 rows with one as its value. The next closest was 2 with 6050.



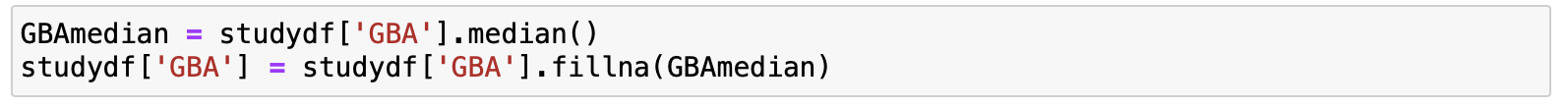




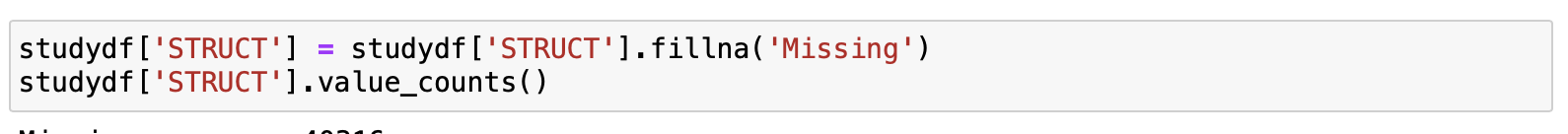




* The values for GBA are skewed to the right. The median was used to fill the null values.



* STRUCT, GRADE, CNDTN, and ROOF did not have a classification that had more than the others. A 'Missing' class was used to fill the null values.

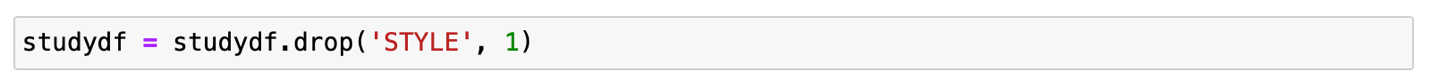








* STYLE is redundant to STORIES. This column was dropped.



* AYB (year built) had 112 rows with null values. These rows were removed.



* AYB and EYB (year of improvement) are float and int types, respectively. These features were changed to DateTime with the year only. (Code from: https://stackoverflow.com/questions/66468687/how-can-i-convert-float-column-to-datetime)

A picture containing logo

Description automatically generated

* The summary statistics and boxplots were used to find outliers in the features. Several of the features have values that are not realistic. For example, the maximum value of STORIES is 826. It is not realistic for a house to have that many stories. After inspecting the boxplots, the values that were above average were removed.

Table

Description automatically generated

(Code From: https://stackoverflow.com/questions/51777217/how-to-plot-a-boxplot-for-each-column-in-a-dataframe)

A picture containing diagram

Description automatically generated

Chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart

Description automatically generatedChart, scatter chart

Description automatically generatedChart

Description automatically generatedA picture containing chart

Description automatically generated

Chart, box and whisker chart

Description automatically generatedChart

Description automatically generatedChart, scatter chart

Description automatically generated

Text

Description automatically generated

* Duplicates were dropped.



* A heatmap was created to examine the correlation between features. Kitchen and the number of units are strongly correlated. Rooms/Bedrooms and AYB/EYB have a high correlation.

Chart

Description automatically generated with medium confidence

* Dummy variables were created for categorical features.

Text

Description automatically generated

* The prepared data set was saved as a csv file.

Analysis

In this study, Ordinary Least Squares (OLS) regression was used to find the relationship between Price and the predictive features. OLS does this by finding coefficients for the predictive features that minimize the sum of the squares of the differences between the observed Price and the predictive Price. (Sharma, 2020) The step to perform an OLS regression are as follows:

* Split the data set into Price and the predictive features.



* Add a constant term. Statsmodels does not include a constant and must be added. This constant functions as the intercept in the model.



* Train and fit the model.



* Run a summary of the results



Table

Description automatically generated

Text, table

Description automatically generated

Text

Description automatically generated

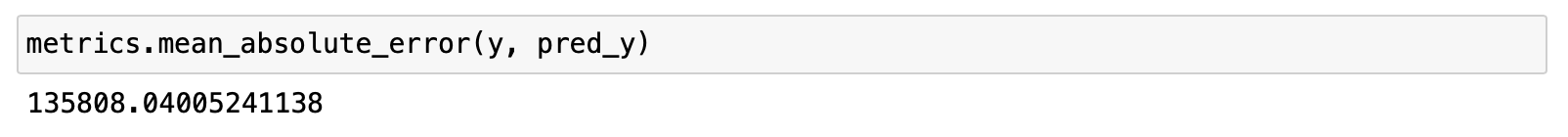
Text

Description automatically generated with medium confidence

Text, table

Description automatically generated

* The mean absolute error was calculated.



* A residual plot was created.

Chart

Description automatically generated

Ordinary Least Squares was chosen for this study because the dependent variable is continuous. Another consideration was that the model summary offers the p-values for the predicting features. One advantage of OLS is it is efficient. It can create models of large data sets without needing extra computing power. One disadvantage is that OLS does not handle outliers well. Since the emphasis of OLS is to minimize the sum of the errors, any value that differs from the rest of the values by a significant amount will have a disproportionately large effect on the model. (ClockBackward, n.d.)

Results

The model has an R-squared value of 44.5%. The model explains 44.5% of the variance. The mean absolute error is $135808, which means that the predicted price is off by this amount on average. The residuals plot does appear to have an equal number of points above and below the line. It is hard to tell if they are spread evenly because of the density of the points themselves.

The main result of this study is that the null hypothesis is rejected. Ward was found to be statistically significant. Of the 105 features used, 37 were found to be statistically significant.

They are as follows:

|  |  |
| --- | --- |
| Feature | Coefficient |
| GRADE\_Good\_Quality | 802500 |
| INTWALL\_Terrazo | 351600 |
| GRADE\_Exceptional-B | 289800 |
| ROOF\_Wood-FS | 197900 |
| INTWALL\_Parquet | 197000 |
| INTWALL\_Vinyl Comp | 158300 |
| INTWALL\_Vinyl Sheet | 130400 |
| CNDTN\_Very Good | 128300 |
| CNDTN\_Excellent | 121300 |
| GRADE\_Very Good | 116200 |
| GRADE\_Superior | 102000 |
| GRADE\_Exceptional-A | 101700 |
| INTWALL\_Default | 87630 |
| ROOF\_Neopren | 85900 |
| BATHRM | 85040 |
| CNDTN\_Good | 71200 |
| CNDTN\_Poor | 67000 |
| INTWALL\_Hardwood | 65770 |
| INTWALL\_Wood Floor | 48350 |
| HF\_BATHRM | 44944 |
| FIREPLACES | 38140 |
| ROOF\_Comp Shingle | 36620 |
| INTWALL\_Hardwood.Carp | 27860 |
| BEDRM | 23380 |
| KITCHENS | 14790 |
| ROOMS | 3189.8637 |
| EYB | 2259.1784 |
| LANDAREA | -2.9952 |
| GBA | -53.5925 |
| AYB | -1436.2863 |
| ROOF\_Shake | -24230 |
| WARD | -24850 |
| STORIES | -25050 |
| ROOF\_Clay Tile | -48410 |
| NUM\_UNITS | -64960 |
| ROOF\_Composition Ro | -75970 |
| AC\_N | -84690 |

One limitation of this study is the amount of data that is missing. Deleting rows and adding a missing class to some features may have introduced bias into the study. The new 'Missing' class also added new features to the study when the categorical features were encoded, which may have contributed to the model's poor performance.

Based on this study, we can see that the feature that adds the most value to a home is having a grade of good. Not having an air conditioner lowers the home price the most. While all these features are significant to the price of a house, investors and homebuyers should focus on the features that add the most value, like condition, grade, and the type of interior.

One direction that a future study could take is to take a non-linear approach to the data. The data does not meet all the assumptions of using a linear model. A non-linear model may be better at predicting the final home price.

A second direction for a future study would be to see if the sales price has a systematic pattern over a period of time.

# Works Cited

Chrisc. (2018). *D.C. Residential Properties*. Retrieved March 2022, from Kaggle: https://www.kaggle.com/datasets/christophercorrea/dc-residential-properties?select=DC\_Properties.csv

ClockBackward. (n.d.). *Ordinary Least Squares Linear Regression: Flaws, Problems and Pitfalls*. Retrieved May 2022, from University of Illinois Chicago: https://lcn.people.uic.edu/classes/che205s17/docs/che205s17\_reading\_06c.pdf

Kishan, H., & Ganguly, S. (2022, March 2). *U.S. house prices to rise another 10% this year*. Retrieved May 2022, from Reuters: https://www.reuters.com/business/us-house-prices-rise-another-10-this-year-2022-03-02/

Orton, K. (2022, April 28). *2021 was another strong year for the D.C. area housing market*. Retrieved May 2022, from The Washington Post: https://www.washingtonpost.com/business/2022/04/28/2021-dc-area-housing-market/

*Python Advantages and Disadvantages - Step in the right direction*. (n.d.). Retrieved March 2022, from TechVidvan.com: https://techvidvan.com/tutorials/python-advantages-and-disadvantages/

Sharma, N. (2020, August 13). *Ordinary Least Squared (OLS) Regression*. Retrieved May 2022, from Medium: https://medium.com/analytics-vidhya/ordinary-least-squared-ols-regression-90942a2fdad5