Executive Summary

Problem Statement and Hypothesis

Problem Statement: Home prices in the United States have risen substantially over the past two years. According to Reuters, the average home price has risen 17% over the last year and is expected to increase 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 asking prices for all properties. (Orton, 2022) This study will use Ordinary Least Squares (OLS) regression to explore the home's features and find if these features affect the home's price.

Research Question and Hypothesis: Does ward affect the price of a home in Washington D.C.? What other features affect the price? The null hypothesis is 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 Analysis Process

Data Collection: The data set for this study was downloaded from D.C. Residential Properties on Kaggle, where it is publicly available. (Chrisc, 2018) It contains 159,000 records of homes in Washington, D.C. The data set has 49 columns. Twenty-two of these columns were used in this study. The columns used are:

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 the main portion of the house was built
EYB	Numeric	Year of the 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

Data Preparation: Python was the primary tool used in this study. Pandas and NumPy were used to import, clean, and transform the data set. Seaborn and Matplotlib were used to create visualizations. The data was loaded into a Jupyter notebook for cleaning and analysis. A new DataFrame was made with the columns listed above. All rows where PRICE was null were removed. The word 'ward' was removed from all entries in the WARD column, and the data type was changed to an integer.

The percent of nulls was calculated for each column. There were 11 columns with 25% of the data missing. The mode was used to replace the missing data for NUM_UNITS, STORIES, EXTWALL, INTWALL, and KITCHENS. The median was used to fill the null values for GBA. A 'Missing' class was created to replace nulls in the STRUCT, GRADE, CNDTN, and ROOF columns. STYLE was found to be redundant to STORIES and was dropped. AYB has 112 rows with null values that were removed. AYB and EYB were changed to DATETIME.

Summary statistics and boxplots were used to find outliers. Outliers were removed for the following columns: PRICE, BATHRM, HF_BATHRM, STORIES, KITCHENS, FIREPLACES, ROOMS, BEDRM, LANDAREA. All duplicates were dropped. Dummy variables were created for the categorical features.

Analysis: Ordinary Least Squares regression was used to find the relationship between PRICE and the predictive features. The data set was split into PRICE and all other columns. A constant term was added to serve as the intercept. The model was trained and fitted. A summary of the model was run.

Results: The model has an R-squared value of 44.5%. The mean absolute error is $135808. A residual plot was created. The distribution of points appears to be even, but it is hard to read, given the density of the points.

The main result of the study is that the null hypothesis is rejected. Ward is statistically significant. Thirty-seven other features were also 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

Limitations of the Technique and Tools

A limitation of this study is the amount of missing data. 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 analysis when the categorical features were encoded, contributing to the model's poor performance.

Proposed Actions

This study shows that the feature that adds the most value to a home is a good grade. Not having an air conditioner has the most significant adverse effect on a home's price. Investors and homebuyers should focus on the features that add the most value, like condition, grade, and the type of interior.

Future studies could take a non-linear approach to model the data to see if a better prediction model could be built. Another method could be exploring if the sales price has a systematic pattern over a period of time.

Expected Benefits

The expected benefit of this study is that investors and homebuyers will know what features of a home affect the price and by how much. New Investors or first-time homebuyers in Washington D.C. could benefit from this study. Since they have no experience in real estate, this study will help them understand which home features are essential to the final price of a home. They can use this information to ensure they are not overpaying for a home. If two houses have all the same features, but one has an air conditioner, they will know the price for the home with an AC should be more. Buyers interested in flipping houses can use this study to see which features to invest money in to get a larger return.

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

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/