



 **maureen-wk / dsc-phase-2-project-v2-3** Public


forked from [learn-co-curriculum/dsc-phase-2-project-v2-3](#)


 View license


 **0** stars


 **367** forks


 Activity


 Star


 Watch


 **Code**


 Pull requests


 Actions


 Projects


 Wiki

 Security


 Insights


 Settings


 main ▾





This branch is **2 commits ahead** of learn-co-curriculum:main.



 Contribute ▾

 Sync fork ▾

 **maureen-wk** final commit ...

2 minutes ago  **18**

 View code

 **README.md** 

Phase 2 Project

Final Project Submission

Please fill out:

- Student name: GROUP 7
- Members: Maureen Kariuki, Aisha Mbarak, MariaCharlotte Mbiyu, Jared Kiprotich, Lee Kamaita, Samuel Lumumba, Edward Opollo
- Student pace: part time
- Scheduled project review date/time:
- Instructor name:
- Blog post URL:

Business Problem

To initiate the project, the following business problems have been formulated for analysis:

Q1. To determine Property Valuation by considering the impact of various property attributes

Q2. To identify the most influential features in determining property prices

Q3. To evaluate potential real estate investment opportunities thus assessing profitability and potential ROI

Project Overview

As research consultants, our objective is to provide valuable insights and comprehensive information to support our stakeholder: **The National Association of Realtors (NAR)**, in advising their clients, including homeowners and property owners, about the impact of various factors on home sale prices in the county.

The project primarily employs multiple linear regression modeling to analyze house sales in a northwestern county.

The outcomes of this project will yield actionable insights that can greatly benefit members of the NAR in the following ways:

1. Facilitating sales growth: The insights gained from the model will help identify key factors influencing home sale prices, enabling NAR members to develop strategies to enhance sales performance.

2. Informing policy implementation: By understanding how different factors impact home prices, NAR can implement effective policies that support homeowners and promote a healthy real estate market.
3. Ensuring long-term customer satisfaction: The insights obtained will enable NAR members to provide informed guidance to homeowners, ensuring their satisfaction and long-term success in real estate transactions.

Ultimately, the model created through this project will empower property buyers and sellers to make well-informed decisions by considering the various factors influencing home sale prices.

Data Understanding

This project uses the King County House Sales dataset, which can be found in `kc_house_data.csv` which is part of this submission. The data contains information about house sales in a northwestern county.

It includes the below features to name a few :

- price
- bedrooms
- bathrooms
- sqft_living
- Zipcode
- Yr built

Short Explanation on the data.

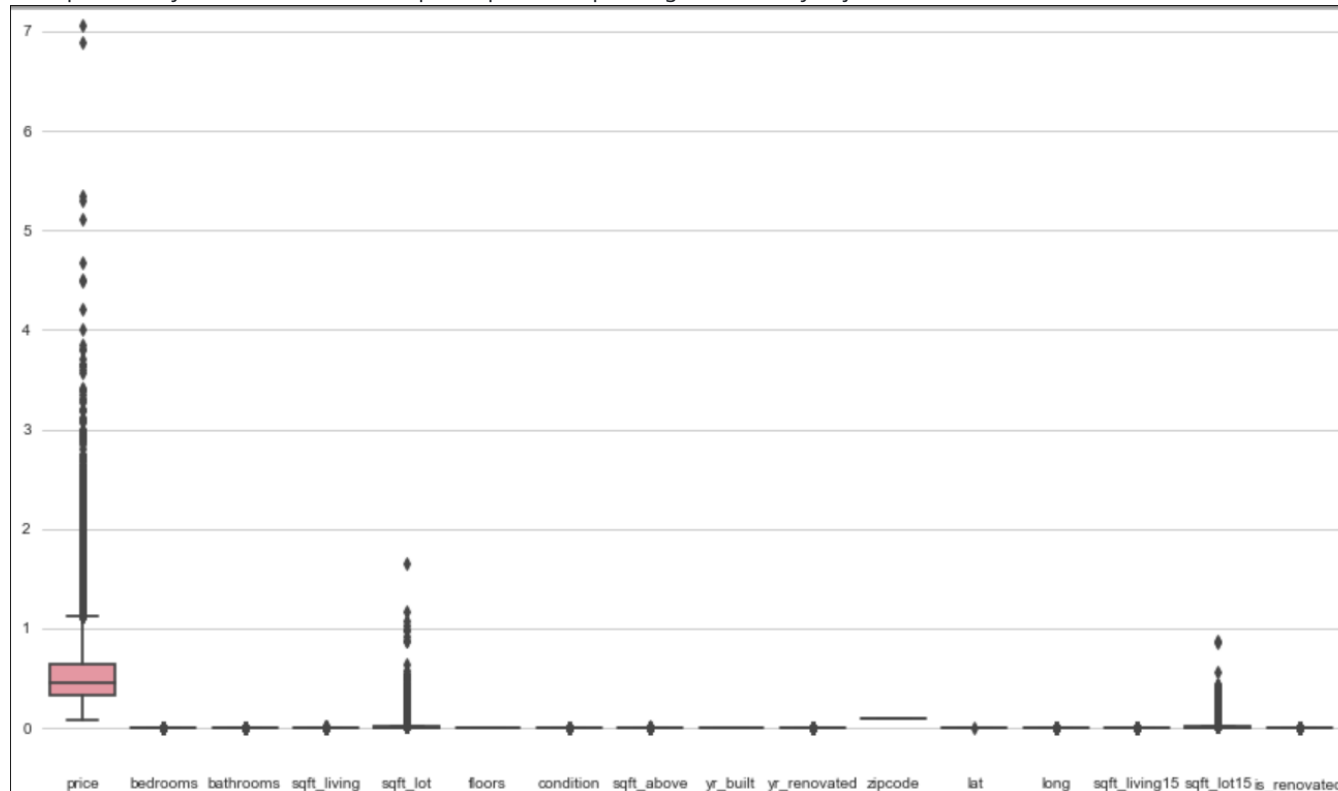
- This is a Pandas Dataframe with 21597 rows and 21 columns.
- The data types in the data frame are 6 floats, 9 intergers (both numerical figures) and 6 objects(categorical figures)
- Missing values can be identified by taking number of entries minus the non null count per column.
- The available columns are as follows: 'id', 'date', 'price', 'bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'waterfront', 'view', 'condition', 'grade', 'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'zipcode', 'lat', 'long', 'sqft_living15', 'sqft_lot15'
- The Memory usage for this dataframe is 3.5+ KB

Data Cleaning and Preparation

To clean the data in preparation for analysis, we start with :

1. Check duplicates in the 'id' column.
2. Drop duplicates if necessary.
3. Identify and handle NAN (Not a Number) /missing values.
4. Check for place holders in 'price'column i.e 0.00
5. Convert data date types if necessary.
6. Identify outliers and either drop / keep them depending on the study objective.
7. Feature Engineering by creating new columns ie 'is_renovated'.
8. Determining columns that are irrelevant for the analysis and drop them.

Example Identify outliers and either drop / keep them depending on the study objective.



Before dropping outliers: (17616, 21) After dropping outliers: (17607, 21)

short explanation of the cleaned dataframe

- The cleaned DataFrame has 17,608 rows and 12 columns.
- The columns are 'date', 'price', 'bedrooms', 'bathrooms', 'sqft_living', 'condition', 'grade', 'sqft_above', 'yr_built', 'yr_renovated', 'sqft_living15', 'is_renovated'.
- The 'date' column has a datetime64 data type.
- The 'price', 'bathrooms', 'yr_renovated', and 'grade' columns have float64 data type.
- The 'bedrooms', 'sqft_living', 'condition', 'sqft_above', 'yr_built', 'sqft_living15', 'is_renovated' columns have int64 data type.
- The total memory usage of the DataFrame is approximately 1.7+ MB.

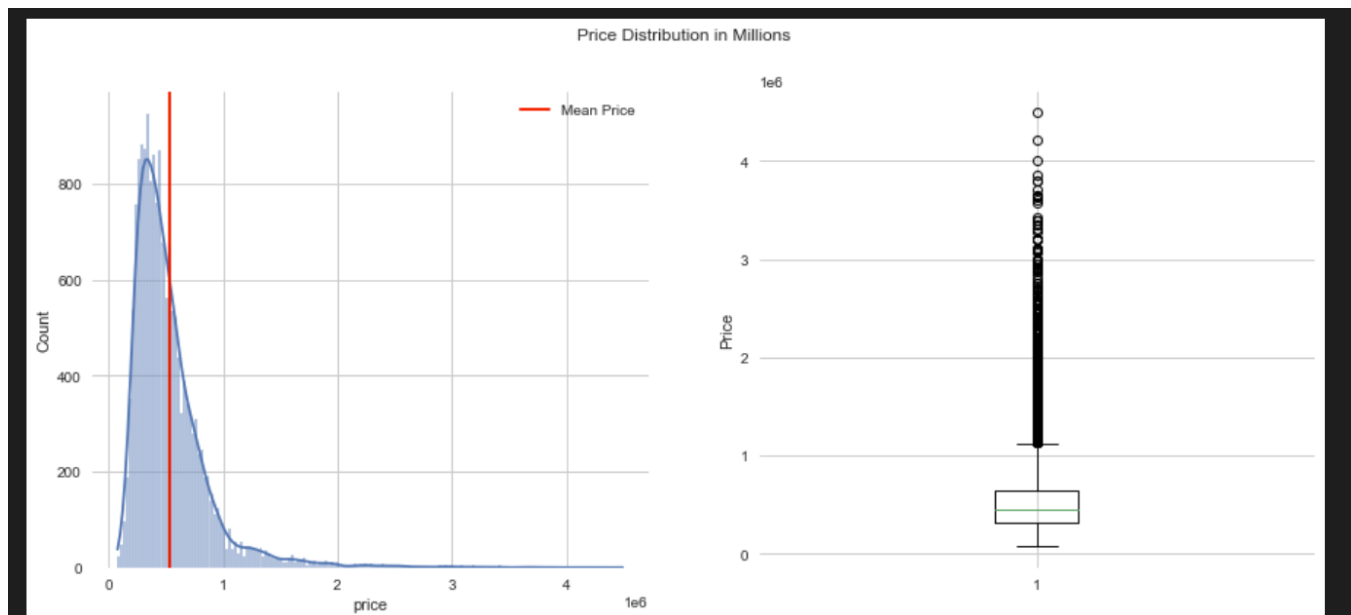
Exploratory Data Analysis

In this step we perform statistical and visualization techniques in order to uncover patterns, relationships, and insights within the data.

- Both Univariate and Bivariate analysis are covered in this section.
- We utilise `df.describe()` and also visualise the columns.
- The output gives a good idea of the central tendency, variability and range of the variable we are looking into.

The analysis is done on 5 columns

- Price
- Bedrooms
- Bathrooms
- sqft_Living
- Grade
- Condition



Summary of Univariate Analysis:

1. Bedrooms:

- On average, the houses in the dataset have approximately 3.4 bedrooms.
- The house with the fewest bedrooms in the dataset has 1 bedroom.
- Most houses have either 3 or 4 bedrooms.
- The house with the most bedrooms in the dataset has 11 bedrooms.

2. Bathrooms:

- On average, the houses in the dataset have around 2.12 bathrooms.
- The house with the fewest bathrooms in the dataset has 0.5 bathrooms.
- Most houses have either 1.75, 2.25, or 2.5 bathrooms.
- The house with the most bathrooms in the dataset has 8 bathrooms.

3. Living Area:

- The average size of the living area in the houses is about 2,083.45 square feet.
- The house with the smallest living area in the dataset is 370 square feet and the largest living area is 13,540 square feet.

4. Condition:

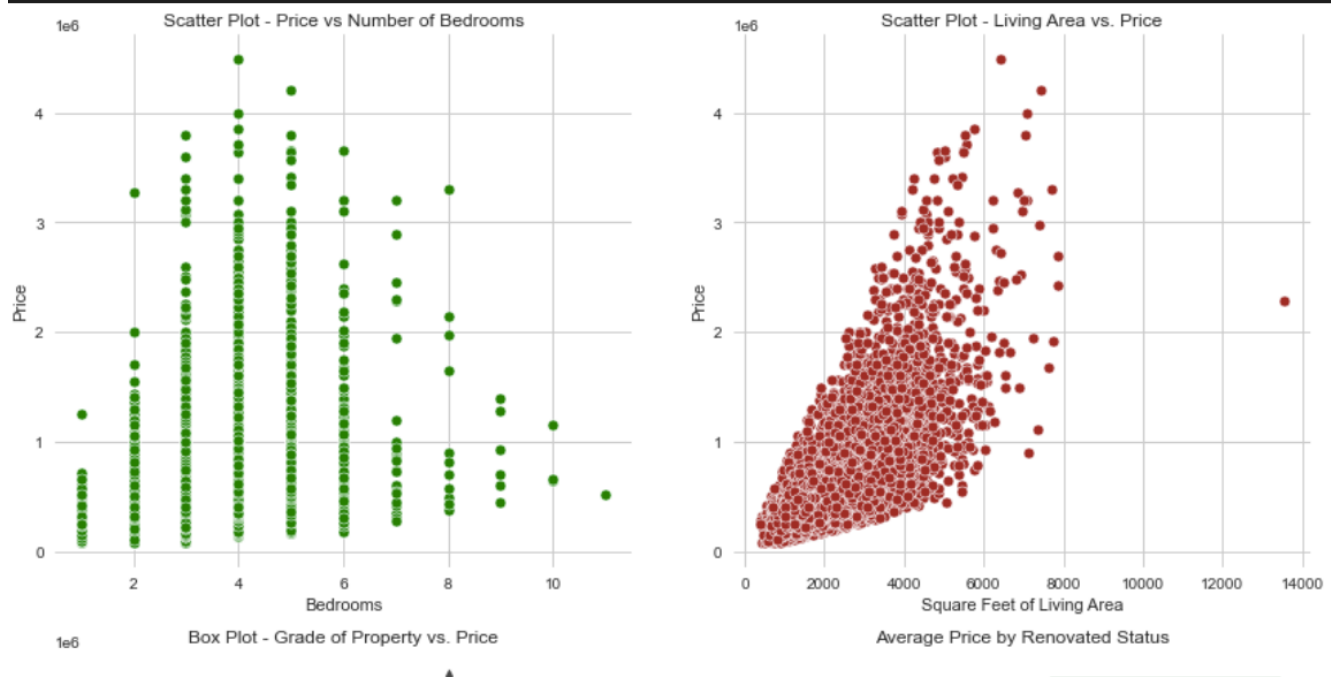
- On average, the houses have a condition rating of 3.41, which indicates the overall state of the house.
- The lowest condition rating in the dataset has a rating of 1, which indicates a poorer condition.
- it should be noted, most houses have a condition rating of either 3 or 4.
- The house with the highest condition rating in the dataset has a rating of 5, which indicates a better condition.

5. Above Ground Living Area:

- The average size of the above ground living area is about 1,791.59 square feet.
- The house with the smallest above ground living area in the dataset is 370 square feet and largest is 9810 square feet.

6. Yr Built:

- On average, the houses in the dataset were built around the year 1971.
- The oldest house in the dataset was built in the year 1900 while the most recent house was built in the year 2015



Summary of Bivariate Analysis

The provided analysis indicates a clear linear correlation between the price (target variable) and several independent variables.

The independent variables considered in this analysis are as follows:

- Number of bedrooms
- Living area space
- Square footage of living space (sqft_living)
- Property grade
- Renovation status

Relationship between Bedrooms and Price: A positive linear relationship is evident, indicating that houses with more bedrooms tend to be more expensive. However, after reaching 7 bedrooms, the price starts to decrease.

Relationship between Living Area Space and Price: The cost of a house generally increases with a larger living area. However, there are instances where houses with large living spaces are priced lower, which could be influenced by other factors.

Relationship between Condition and Price: The condition of a house affects its pricing. Houses in average to very good condition tend to have higher prices.

Relationship between Grade and Price: A positive linear relationship exists between the grade of a property and its price. This is particularly noticeable for poorly and low-graded houses, which typically have lower prices.

Relationship between Renovation and Price: There is a positive correlation between houses that have been renovated and higher prices.

A house that possesses most of the above variables will command a higher price in the market, while houses with weaker performance in these variables will be comparatively cheaper.

Model Creation

Simple Linear Regression

Model 1 :Creating a Baseline

OLS Regression Results

```

=====
Dep. Variable:          price      R-squared:                0.485
Model:                  OLS        Adj. R-squared:           0.485
Method:                 Least Squares    F-statistic:           1.658e+04
Date:                  Fri, 02 Jun 2023    Prob (F-statistic):      0.00
Time:                  17:35:08      Log-Likelihood:        -2.4398e+05
No. Observations:      17607        AIC:                   4.880e+05
Df Residuals:          17605        BIC:                   4.880e+05
Df Model:               1
Covariance Type:       nonrobust
=====
               coef      std err          t      P>|t|      [0.025      0.975]
-----
const          -2.425e+04    4767.056      -5.086      0.000     -3.36e+04     -1.49e+04
sqft_living     270.1629         2.098     128.757      0.000      266.050      274.276
=====
Omnibus:                9589.607    Durbin-Watson:           1.966
Prob(Omnibus):           0.000    Jarque-Bera (JB):        147439.407
Skew:                    2.281    Prob(JB):                 0.00
Kurtosis:                16.422    Cond. No.                 5.70e+03
=====

```

Model 1: Simple Linear Regression Results

Looking at the summary above, the regression line we found is

$$\hat{price} = -24,220 + 270.16sqft_{living}$$

- Our y intercept in Model 1 is -24,220.
- The model is statistically significant, with an F-statistic p-value well below 0.05
- The model (R-squared) explains about 48.5% of the variance in price.
- The model coefficients (const and sqft_living) are both statistically significant, with t-statistic p-values well below 0.05
- If a house has sqft_living space of 0 feet squared, we would expect the price to be about USD -24,220
- For each increase of 1 square foot in sqft_living space, the price increases by USD 270.16

Multiple linear regression

Model 2: Columns with correlation >50% with 'price'

OLS Regression Results

=====						
Dep. Variable:	price	R-squared:	0.495			
Model:	OLS	Adj. R-squared:	0.495			
Method:	Least Squares	F-statistic:	5760.			
Date:	Fri, 02 Jun 2023	Prob (F-statistic):	0.00			
Time:	17:35:11	Log-Likelihood:	-2.4380e+05			
No. Observations:	17607	AIC:	4.876e+05			
Df Residuals:	17603	BIC:	4.876e+05			
Df Model:	3					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	-8.744e+04	6539.196	-13.371	0.000	-1e+05	-7.46e+04
bathrooms	-1084.4037	3733.935	-0.290	0.771	-8403.286	6234.479
sqft_living	224.6080	3.982	56.412	0.000	216.804	232.412
sqft_living15	80.5831	4.234	19.031	0.000	72.283	88.883
=====						
Omnibus:	9835.975	Durbin-Watson:	1.967			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	158715.082			
Skew:	2.345	Prob(JB):	0.00			
Kurtosis:	16.941	Cond. No.	1.11e+04			
=====						

Model 2 Results:

The second Model built illustrates price as below:

$$\hat{price} = -87,440 - 1089.08bathrooms + 224.61squarefootliving + 80.55sqftliving15$$

- Our y intercept in this model is -87,440
- The model is statistically significant overall, with an F-statistic p-value well below 0.05
- The model explains approximately 49.5% of the variability in the dependent variable (price)
- This is a 1% increase from our baseline model and thus may not have much of a difference.
- The model coefficients (const , sqft_living and sqft_living15) are all statistically significant, with t-statistic p-values way below 0.05.
- However, the bathroom coefficient is not statistically significant. We can thus drop it for our next model.
- On average, each additional square foot of living area is associated with an increase of approximately USD224.61 in the price.
- This is a decrease of approximately 45 dollars from the baseline model. This may mean that the additional variables have significance in the relationship between sqft_living and price.
- For each increase of 1 square foot living15 in a house , there is an associated price increase of USD 80.58

The Partial regression plot displays the data above and is consistent with the model findings.

Overall, this regression model suggests that the number of bathrooms has no significant effect on the price, while the square footage of the living area and the square footage of the neighboring properties' living area have significant positive effects on the price.

Model 3: All correlated columns minus bathrooms

We create a multiple linear regression by utilising all columns with the positively correlated predictors.

We will exclude Bathrooms from this model as it is not statistically significant as per model 2.

https://github.com/Kelta153/dsc_phase-2_project/blob/main/images/m3.png

OLS Regression Results

=====						
Dep. Variable:	price	R-squared:	0.550			
Model:	OLS	Adj. R-squared:	0.550			
Method:	Least Squares	F-statistic:	3592.			
Date:	Fri, 02 Jun 2023	Prob (F-statistic):	0.00			
Time:	17:35:13	Log-Likelihood:	-2.4278e+05			
No. Observations:	17607	AIC:	4.856e+05			
Df Residuals:	17600	BIC:	4.856e+05			
Df Model:	6					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	4.208e+06	1.45e+05	28.966	0.000	3.92e+06	4.49e+06
bedrooms	-5.735e+04	2460.576	-23.308	0.000	-6.22e+04	-5.25e+04
sqft_living	272.4615	3.495	77.951	0.000	265.610	279.313
condition	2.035e+04	2985.938	6.816	0.000	1.45e+04	2.62e+04
yr_built	-2184.0993	72.242	-30.233	0.000	-2325.700	-2042.498
is_renovated	9.443e+04	9325.958	10.125	0.000	7.61e+04	1.13e+05
sqft_living15	94.5853	4.049	23.363	0.000	86.650	102.521
=====						
Omnibus:	9602.158	Durbin-Watson:	1.962			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	165242.016			
Skew:	2.245	Prob(JB):	0.00			
Kurtosis:	17.321	Cond. No.	2.95e+05			
=====						

Model 3 Results:

The third Model built illustrates price as below:

$$\hat{price} = 4,208,000 - 57,350bedrooms + 272.46squarefootliving + 20,350condition - 2180.09yrbuilt + 94,430isrenovated + 94.58sqftliving15$$

- Our y intercept in this model is 4,208,000
- The model is statistically significant with an F-statistic p-value well below 0.05
- The model explains approximately 55% of the variability in the dependent variable (price)
- The model coefficients (const , bedrooms , sqft_living , condition , yr_built , is_renovated and sqft_living15 are all statistically significant, with t-statistic p-values well below 0.05.
- On average, each additional bedroom is associated with a decrease of approximately USD 57,350 in the price.
- For each additional square foot of living area is associated with an increase of approximately USD272.46 in the price.
- This is a decrease of USD 2.3 from our baseline model and an increase of USD 48 from our second model.
- On average, each unit increase in condition is associated with an increase of approximately USD20,350 in the price.
- The yr_built on the other hand has an associated decrease in price the older the house becomes by approximately USD 2184
- A renovated property increases the price by USD 94,300
- On average, each additional square foot of the neighboring properties' living area is associated with an increase of approximately USD 94.59 in the price.

Model 4: Log Transformed data

For this model, we log transformed our data to improve our final model.

OLS Regression Results

=====						
Dep. Variable:	price	R-squared:	0.458			
Model:	OLS	Adj. R-squared:	0.458			
Method:	Least Squares	F-statistic:	2480.			
Date:	Fri, 02 Jun 2023	Prob (F-statistic):	0.00			
Time:	17:35:20	Log-Likelihood:	-2.4443e+05			
No. Observations:	17607	AIC:	4.889e+05			
Df Residuals:	17600	BIC:	4.889e+05			
Df Model:	6					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	2.81e+07	1.17e+06	23.979	0.000	2.58e+07	3.04e+07
bedrooms	-2.016e+05	9178.617	-21.960	0.000	-2.2e+05	-1.84e+05
sqft_living	5.157e+05	8530.382	60.450	0.000	4.99e+05	5.32e+05
condition	5.466e+04	1.17e+04	4.677	0.000	3.18e+04	7.76e+04
yr_built	-4.351e+06	1.56e+05	-27.850	0.000	-4.66e+06	-4.04e+06
sqft_living15	2.408e+05	9142.891	26.342	0.000	2.23e+05	2.59e+05
is_renovated	4313.6054	445.242	9.688	0.000	3440.888	5186.323
=====						
Omnibus:	11754.486	Durbin-Watson:	1.953			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	281587.942			
Skew:	2.862	Prob(JB):	0.00			
Kurtosis:	21.737	Cond. No.	1.58e+04			
=====						

Model 4 Results:

The log transformed variables do not improve the fit of the model compared to model 3.

This can be attributed to the Zeros in 'is_renovated' column which needed to be added a small epsilon value. The third Model built illustrates price as below:

$$\hat{price} = 28,100,000 - 201,600bedrooms + 515,700squarefootliving + 54,660condition - 4,351,000yrbuilt + 4313isrenovated + 240,800sqftliving15$$

- Our y intercept in this model is \$28,100,000
- The model is statistically significant with an F-statistic p-value well below 0.05
- The model explains approximately 45.8% of the variability in the dependent variable (price)
- The model coefficients are all statistically significant, with t-statistic p-values well below 0.05.

Chosen Model: Model 3

After evaluating the 4 models created, we settled on the Model 3 because :

1. With the highest R-squared value of 55%, our third Model outperforms the other models in explaining the majority of the variability in price. This indicates a better fit for the data while avoiding overfitting.
2. Model 3 exhibits the lowest Mean Absolute Error, approximately 158,420. This implies that the predictions made by this model have the smallest overall deviation from the actual values, regardless of the direction of the deviation. It thus demonstrates better accuracy and performance.
3. Model 3 incorporates the most features from the dataframe, with only one feature being deemed statistically insignificant and excluded from the model. This suggests that Model 3 takes into account a comprehensive set of variables, potentially capturing more nuances and improving the predictive accuracy.

Conclusion

The following conclusions were drawn from this project and in the process answering the 3 business problems stated earlier

To determine Property Valuation by considering the impact of various property attributes

To evaluate potential real estate investment opportunities thus assessing profitability and potential ROI

1. The model shows a moderate level of predictive power. The R-squared value of 0.550 indicates that the independent variables included in the model can explain approximately 55% of the variability in home prices. This suggests that the selected features have some influence on the pricing of homes.

The below is the property valuation model: $\hat{\text{price}} = 4,208,000 - 57,350 \text{ bedrooms} + 272.46 \text{ squarefootliving} + 20,350 \text{ condition} - 2180.09 \text{ yrbuilt} + 94,430 \text{ isrenovated} + 94.58 \text{ sqftliving15}$

To identify the most influential features in determining property prices

2. Significant predictors of price as per the model are the number of bedrooms, square footage of living area, condition of house, year built, whether the property has been renovated, and the square footage of neighboring properties. These variables demonstrate a significant association with the dependent variable, indicating their importance in determining the price of a home.
3. Normality assumption: The Q-Q plots of the model's residuals suggest that they approximately follow a normal distribution. This indicates that the assumption of normality is reasonably met, which is important for the validity of the statistical inference and interpretation of the model results.

In summary, the study suggests that the number of bedrooms, square footage, condition, year built, renovations, and neighboring property characteristics are important factors to consider when determining the price of a home. However, it is essential to consider other market factors and property-specific attributes in conjunction with the findings of this analysis to arrive at an accurate and competitive listing price for example availability of different amenities such as schools, shopping malls, hospitals, factories etc.

Recommendations

Recommendations to Homeowners

Based on the findings from the regression analysis, the following recommendations can be made to homeowners:

1. Consider the number of bedrooms: The coefficient for the "bedrooms" variable is negative, indicating that an increase in the number of bedrooms may have a negative impact on the house price. Homeowners should carefully evaluate their needs and the market demand for different bedroom configurations when making decisions about the number of bedrooms in their homes.
2. Focus on the square footage: The coefficient for "sqft_living" suggests that an increase in square footage positively influences the house price. Homeowners should consider investing in home improvements or expansions that increase the living space, as it may have a positive impact on the value of their property.
3. Maintain the condition of the property: The coefficient for the "condition" variable indicates that a higher condition rating positively affects the house price. Homeowners should prioritize regular maintenance and repairs to keep their homes in good condition, which can potentially enhance the market value.
4. Pay attention to the year built: The coefficient for "yr_built" suggests that older homes may have a negative impact on the price. Homeowners of older properties could consider renovations or updates to modernize their homes and potentially increase their market value.
5. Renovations can add value: The coefficient for the "is_renovated" variable indicates that homes that have been renovated have a positive impact on the price. Homeowners who are considering renovations should carefully plan and budget for these improvements, as they can potentially yield a higher return on investment.
6. Consider the influence of neighboring properties: The coefficient for "sqft_living15" suggests that the square footage of nearby properties (within a certain radius) can influence the house price. Homeowners should be aware of the market trends and the characteristics of neighboring properties, as these factors can impact the value of their own homes.

Overall, homeowners should consider these factors but also consult with real estate professionals for a more comprehensive analysis tailored to their specific property and market conditions.

Recommendations to Members NAR

As members of the National Association of Realtors, real estate professionals play a crucial role in guiding their clients through the buying and selling process. Based on the findings from the regression analysis, here are some recommendations for members of the National Association of Realtors:

1. Stay updated on market trends: Continuously monitor and analyze market trends, including factors such as the number of bedrooms, square footage, property condition, year built, renovations, and neighboring property characteristics. This information will help you provide accurate and valuable insights to your clients.
2. Educate clients on the impact of features: Clearly explain to clients how various features of a property, such as the number of bedrooms, square footage, and condition, can influence its market value. Help them understand the potential trade-offs and considerations when making decisions about buying or selling a property.
3. Provide renovation recommendations: Offer guidance on renovations or updates that can enhance the value of a property. Advise clients on which improvements are most likely to yield a positive return on investment based on the findings from the regression analysis.
4. Conduct thorough market analyses: Before listing a property, perform a comprehensive market analysis that takes into account the local market conditions, recent sales data, and the specific features of the property. Use this information to set an appropriate listing price and advise clients on the potential selling price range.
5. Collaborate with appraisers: Work closely with professional appraisers to ensure accurate property valuations. Share the regression analysis findings with appraisers to provide additional insights and support the appraisal process.
6. Stay informed about regulations and policies: Stay updated on any regulatory changes or policies that may impact the real estate market. This knowledge will help you provide informed advice to your clients and navigate any legal or policy-related challenges.

By following these recommendations, members of the National Association of Realtors can provide valuable guidance to their clients, assist them in making informed decisions, and maintain professionalism and expertise in the real estate industry.

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Languages

● Jupyter Notebook 100.0%