Pruthvik Patel July 21, 2023

Assignment #3

Area of Interest: Real Estate Market in New Jersey

I am personally drawn to exploring the real estate market in New Jersey and creating a predictive model for three-bedroom home prices. It's a project that really captivates me. As someone with a keen interest in real estate investment, I see immense value in understanding the intricacies and variables that influence housing prices. With the ability to predict the prices of three-bedroom homes accurately, I can make well-informed investment decisions and seize opportunities that others may overlook.

Data Identification and Exploration:

To build a predictive model for the price of three-bedroom homes in New Jersey, I would need to identify and explore relevant data sources. Here are some potential data sources to consider:

- **Real Estate Listings:** Obtain data from real estate listing websites or agencies that provide information about properties, including their features, location, and prices.

- **Historical Sales Data**: Collect historical data on the sale prices of three-bedroom homes in New Jersey. This data can help identify trends and patterns over time.

When developing a predictive model for home prices, it is important to consider three key factors. Firstly, property characteristics play a significant role, including square footage, the number of bedrooms and bathrooms, location (such as neighborhood and proximity to schools or transportation), and the overall property condition. Secondly, economic indicators should be taken into account, encompassing elements like interest rates, employment rates, GDP growth, and demographic information such as population density and income levels. Lastly, market trends are crucial to analyze, including data on housing supply and demand dynamics, average days on the market, and price fluctuations. By incorporating these factors into the model, a more comprehensive understanding of home prices can be achieved, leading to more accurate predictions.

Dependent Variable:
Price of three-bedroom homes

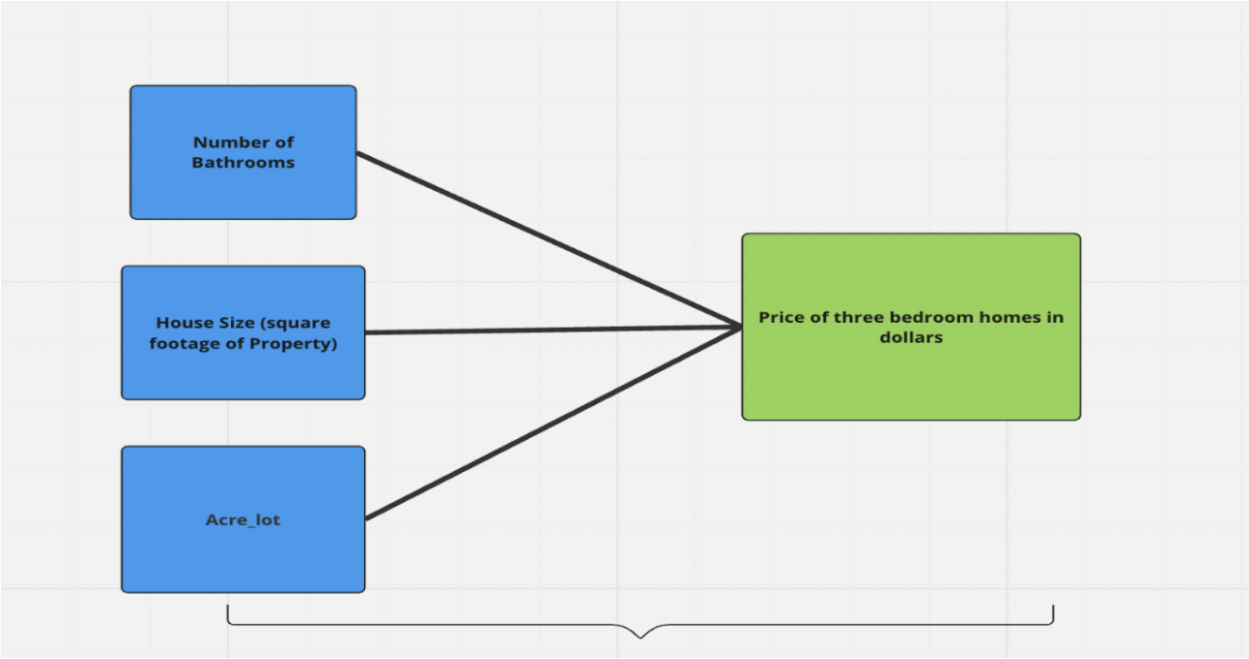
Unit of Analysis:
Each Home

Independent Variables:

Square footage of the property
Number of bathrooms
Lot Size (Number of Acres)

Model and Method for Exploration:

To develop a predictive model for three-bedroom home prices, a suggested framework consists of several stages. Firstly, data collection involves gathering a comprehensive dataset that encompasses the variables of interest, including price, property features, economic indicators, and market trends. This dataset should cover a significant number of three-bedroom homes located across various areas in New Jersey. Secondly, data processing is crucial, which involves cleaning the data by addressing missing values, outliers, and inconsistencies. Additionally, categorical variables need to be converted into numerical representations to facilitate analysis. Moving on to feature selection, it is essential to identify the most influential features that impact the price of three-bedroom homes. Techniques such as correlation analysis and examining feature importance from regression models can aid in this process. Subsequently, the dataset is split into training and testing sets for model training. A multiple linear regression model is constructed, with the price of the home serving as the dependent variable and the selected features as independent variables. Finally, once the model's performance meets the desired criteria, it can be deployed to make predictions on new or unseen data. Continuous monitoring of the model's performance over time is necessary, allowing for refinements to enhance its predictive capabilities as needed.



Unit Of Analysis: Each Home

Challenges and Approaches:

Data Quality and Availability: Ensure that the collected data is reliable, accurate, and representative of the New Jersey real estate market. Address any missing or incomplete data through imputation techniques or data augmentation.

Ethical Considerations: Be mindful of potential biases in the data and ensure fairness and equity in the model's predictions.

Viewpoint: Take a holistic perspective by considering not only property features but also economic indicators and market trends. This broader viewpoint can enhance the accuracy and robustness of the predictive model.

Analytical Approach

Relationships: In developing a predictive model for three-bedroom home prices in New Jersey, I anticipate observing clear relationships between the three selected independent variables and the price of homes. The chosen independent variables are square footage, number of bathrooms, and lot size. Square footage is expected to have a significant impact on home prices. Larger homes with more square footage generally command higher prices compared to smaller homes. This is because more living space is often associated with increased desirability and functionality, resulting in higher market value.

The number of bathrooms is another crucial factor influencing home prices. Homes with more bathrooms tend to be in higher demand and, as a result, can have higher price tags. Additional bathrooms provide convenience and comfort for households, especially in properties accommodating multiple occupants or families.

Lot size can also play a role in determining home prices. A larger lot size can offer potential for outdoor activities, expansion, or privacy, which can contribute to higher property values. Buyers often consider lot size when evaluating properties, and it can be a deciding factor in the perceived worth of a home.

Informed Estimate of Findings: Based on my understanding of the real estate market, I expect that the selected independent variables will demonstrate significant associations with the price of three-bedroom homes in New Jersey. Larger square footage, a greater number of bathrooms, and a larger lot size are likely to be positively correlated with higher home prices.

Homes with more square footage are likely to be priced higher due to the increased livable area they provide. Similarly, properties with a greater number of bathrooms are expected to fetch

higher prices as they offer added convenience and accommodate larger households more comfortably. Larger lot sizes can be considered a desirable feature and may lead to higher property values.

By analyzing the dataset and implementing the multiple linear regression model using these three independent variables, I aim to gain valuable insights into their impact on three-bedroom home prices in New Jersey. These findings will assist in making well-informed investment decisions within the real estate market.

Data Collection:

To gather the data for the real estate market in New Jersey, I utilized a dataset from Kaggle, which was uploaded by Ahmed Sakib and is regularly updated monthly. Kaggle is an online community and platform for data science and machine learning practitioners, providing a wide range of freely available datasets for download and use. The dataset contained over 500,000 rows of data, including real estate pricing for homes across the United States. The dataset included the following columns: housing status (ready for sale or ready to build), number of bedrooms, number of bathrooms, property/land size in acres, city name, state name, postal code of the area, house area/size/living space in square feet, previously sold date, and housing price (either the current listing price or recently sold price if the house was sold recently).

https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset

Data Curation:

During the exploration of the data, I observed the presence of many duplicate rows with identical information. To address this, I utilized the Python Pandas module to remove all the duplicate rows efficiently, as using MS Excel was time-consuming even for highlighting the duplicate values due to the size of the data. Moving forward, I filtered the data specifically for the state of New Jersey and focused on homes with three bedrooms, as my primary interest was predicting the price of three-bedroom homes in New Jersey. Furthermore, when I came across null values in columns like house_size and acre_lot, I had to figure out what to do with them. One option was to fill these empty values with the median value for each column. However, I wasn't sure if this approach would be the best choice. Instead, I thought about finding the average values for homes with certain characteristics, such as the number of bathrooms, and using those averages to replace the missing values. For example, I considered calculating the average square footage for homes with two bathrooms and using that value to fill in the missing data for other homes with two bathrooms. But after giving it more thought, I decided to take a different route. Since I had a good amount of data even after removing the rows with null values, I chose to simply get rid of

those incomplete rows. By doing this, I ensured that the dataset I would use for the linear regression model was complete and didn't require any additional steps. **Below are the snippets of the dataset before and after curation**

Before Curation

	A	B	C	D	E	F	G	H	I	J	K	L
1	status	price	bed	bath	acre_lot	full_addr	street	city	state	zip_code	house_siz	sold_date
2	for_sale	105000	3	2	0.12	Sector Yahue	Sector Yahue	Adjuntas	Puerto Rico	601	920	
3	for_sale	80000	4	2	0.08	Km 78 9 Carr	Km 78 9 Carr	Adjuntas	Puerto Rico	601	1527	
4	for_sale	67000	2	1	0.15	556G 556-G	556G 556-G	Juana Diaz	Puerto Rico	795	748	
5	for_sale	145000	4	2	0.1	R5 Comunida	R5 Comunida	Ponce	Puerto Rico	731	1800	
6	for_sale	65000	6	2	0.05	14 Navarro, I	14 Navarro	Mayaguez	Puerto Rico	680		
7	for_sale	179000	4	3	0.46	Bo Calabaza	Bo Calabaza	San Sebastia	Puerto Rico	612	2520	
8	for_sale	50000	3	1	0.2	49.1 140, Cia	49.1 140	Ciales	Puerto Rico	639	2040	
9	for_sale	71600	3	2	0.08	3467 St, Pon	3467 St	Ponce	Puerto Rico	731	1050	
10	for_sale	100000	2	1	0.09	230 Rio De V	230 Rio De V	Ponce	Puerto Rico	730	1092	
11	for_sale	300000	5	3	7.46	Pr 120 Bo M	Pr 120 Bo M	Las Marias	Puerto Rico	670	5403	
12	for_sale	89000	3	2	13.39	Km 3 4 Solar	Km 3 4 Solar	Isabela	Puerto Rico	662	1106	
13	for_sale	150000	3	2	0.08	91 Del Rio, J	91 Del Rio	Juana Diaz	Puerto Rico	795	1045	
14	for_sale	155000	3	2	0.1	Pr, Lares, PR	Pr	Lares	Puerto Rico	669	4161	
15	for_sale	79000	5	2	0.12	90 # A10, Ut	90 # A10	Utuado	Puerto Rico	641	1620	
16	for_sale	649000	5	5	0.74	F118 Madrid	F118 Madrid	Ponce	Puerto Rico	731	2677	
17	for_sale	120000	3	2	0.08	10-K Alejand	10-K Alejand	Yauco	Puerto Rico	698	1100	
18	for_sale	235000	4	4	0.22	10 Calle Carr	10 Calle Carr	Mayaguez	Puerto Rico	680	3450	
19	for_sale	105000	3	2	0.08	DD18 Calle 2	DD18 Calle 2	Ponce	Puerto Rico	728	1500	
20	for_sale	575000	3	2	3.88	5.8 Carr 435	5.8 Carr 435	San Sebastia	Puerto Rico	685	4000	
21	for_sale	140000	6	3	0.25	1 Bo Corcova	1 Bo Corcova	Anasco	Puerto Rico	610	1230	
22	for_sale	50000	2	1	0.23	Km 5 0 Carr	Km 5 0 Carr	Yauco	Puerto Rico	698	621	
23	for_sale	165000	6	3	0.1	110 Concepc	110 Concepc	Moca	Puerto Rico	676	3000	
24	for_sale	189000	3	1	2	4C Calle Gira	4C Calle Gira	Coamo	Puerto Rico	769	1213	
25	for_sale	115000	3	2		17 Estancias	17 Estancias	Ponce	Puerto Rico	716	1148	
26	for_sale	122500	3	2	0.05	16 Muoz Riv	16 Muoz Riv	Yauco	Puerto Rico	698	1118	
27	for_sale	255000	3	2	0.28	Rd 125 Km 1	Rd 125 Km 1	San Sebastia	Puerto Rico	685	1500	
28	for_sale	425000	4	3	0.3	31 Calle A, P	31 Calle A	Ponce	Puerto Rico	730	3000	
29	for_sale	93000	4	2	0.11	C12 Haciend	C12 Haciend	Guayanilla	Puerto Rico	656	1300	
30	for_sale	75000	4	2	0.04	176 Calle Ne	176 Calle Ne	Manati	Puerto Rico	676	1080	
31	for_sale	469000	3	2	0.69	Carretera 45	Carretera 45	Isabela	Puerto Rico	662	2505	
32	for_sale	189000	6	3	1.75	Km 536 Carr	Km 536 Carr	Villalba	Puerto Rico	766	1943	
33	for_sale	225000	5	4	0.12	Vannina, Por	Vannina	Ponce	Puerto Rico	717	1600	
34	for_sale	495000	4	2	3.08	Sector Usera	Sector Usera	Santa Isabel	Puerto Rico	757	2886	
35	for_sale	220000	1	1	0.1	18 Paseo De	18 Paseo De	Ponce	Puerto Rico	730	850	
36	for_sale	70500	3	1	0.23	Carr 371 Her	Carr 371 Her	Yauco	Puerto Rico	698	936	
37	for_sale	305000	3	2	1.13	Km 6 4 Carr	Km 6 4 Carr	Moca	Puerto Rico	676	1928	
38	for_sale	475000	4	4	4.84	PR-445 Km 3	PR-445 Km 3	San Sebastia	Puerto Rico	685	3690	

After Curation

bath	house_size (sqft)	acre_lot (acre)	price ($)		
3	1500	0.07	333490		
2	1542	0.14	305100		
1	925	0.11	205000		
3	1870	0.15	325000	Unit of Analysis	Each Home
3	1476	0.06	399000	Independent Variable	
2	7501	0.17	340000	Dependent Variable	
1	1008	0.17	325000		
1	1414	0.17	375000		
1	862	0.11	375000		
2	1199	0.11	350000		
2	1232	0.1	449000		
1	1328	0.09	545000		
3	1315	0.06	495000		
3	1232	0.09	385000		
3	2728	0.55	764900		
3	2080	2.48	486000		
3	1901	0.87	499000		
3	1377	0.08	449999		
2	1561	0.28	525000		
4	3799	1.63	1049000		
2	1552	0.19	335000		
2	1104	0.12	1495000		
2	1280	0.08	425000		
2	1500	0.11	459000		
2	1440	0.08	424000		
2	1622	0.17	635000		
2	2032	0.05	449000		

SPSS Statistical Analysis:

Correlations

		price	number_of_bathrooms	house_size_sqft	acre_lot
price	Pearson Correlation	1	.563**	.192**	.164**
	Sig. (2-tailed)		<.001	<.001	<.001
	N	655	655	655	655
number_of_bathrooms	Pearson Correlation	.563**	1	.137**	.088*
	Sig. (2-tailed)	<.001		<.001	.025
	N	655	655	655	655
house_size_sqft	Pearson Correlation	.192**	.137**	1	.049
	Sig. (2-tailed)	<.001	<.001		.208
	N	655	655	655	655
acre_lot	Pearson Correlation	.164**	.088*	.049	1
	Sig. (2-tailed)	<.001	.025	.208	
	N	655	655	655	655

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Correlation Scores

- Number of bathrooms & Home Price → r = 0.563
- House size (sqft) & Home Price → r = 0.192
- Acre lot & Home Price → r = 0.164

Significance Values

- Number of bathrooms & Home Price → $p < 0.001$
    - Statistically significant ($p < 0.05$)
- House size (sqft) & Home Price → $p < 0.001$
    - Statistically significant ($p < 0.05$)
- Acre lot & Home Price → $p < 0.001$
    - Statistically significant ($p < 0.05$)

Correlation Directions

- Number of bathrooms & Home Price → positive direction
- House size (sqft) & Home Price → positive direction

- Acre lot & Home Price → positive direction

Correlations: Interpretations and Analysis

In my regression model aiming to predict the prices of 3-bedroom homes in New Jersey, I examined the correlation scores to assess the relationship between the independent variables and the dependent variable (home price).

The correlation coefficient (r) between the **number of bathrooms and home price** is 0.563, indicating a positive correlation. This implies that as the number of bathrooms increases, the home price tends to increase as well. The relatively high correlation coefficient suggests a moderately strong linear association between these variables.

Similarly, the correlation coefficient **between the house size (sqft) and home price** is 0.192, also indicating a positive correlation. This suggests that as the house size increases, the home price tends to increase. However, the correlation coefficient's smaller value implies a weaker linear association compared to the number of bathrooms.

Furthermore, the correlation coefficient **between the acre lot and home price** is 0.164, once again reflecting a positive correlation. This indicates that as the size of the lot in acres increases, the home price tends to increase as well. However, the correlation coefficient's smaller value suggests a weaker linear association between these variables.

Analyzing the significance values, it is noteworthy that all three independent variables (number of bathrooms, house size, and acre lot) demonstrate statistical significance with p-values less than 0.001. This indicates a high level of confidence in the correlations being genuine and not due to chance.

In summary, based on the correlation analysis of my regression model, the number of bathrooms, house size (sqft), and acre lot are all positively correlated with home prices. However, the strength of the correlations varies, with the number of bathrooms exhibiting a stronger association than the other two variables. Additionally, the statistical significance of all three variables further supports the credibility of their correlations with home prices.

Multiple Linear regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.585[a]	.342	.339	155695.959

a. Predictors: (Constant), acre_lot, house_size_sqft, number_of_bathrooms

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.220E+12	3	2.740E+12	113.025	<.001[b]
	Residual	1.578E+13	651	24241231780		
	Total	2.400E+13	654			

a. Dependent Variable: price

b. Predictors: (Constant), acre_lot, house_size_sqft, number_of_bathrooms

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	170452.975	17732.004		9.613	<.001
	number_of_bathrooms	123090.197	7370.149	.538	16.701	<.001
	house_size_sqft	9.265	2.642	.113	3.507	<.001
	acre_lot	22470.156	6438.179	.111	3.490	<.001

a. Dependent Variable: price

Regression Summary:

The regression analysis yielded an R-square value of 0.342. This indicates that approximately 34.2% of the variance in home prices can be explained by the independent variables included in the model. It is worth noting that the R-square value is relatively moderate, suggesting that there may be other factors beyond the included variables that influence home prices in New Jersey.

Significant Variables & Coefficients:

The significance of each independent variable and their coefficients were evaluated to assess their predictive relationship with home prices.

Number of Bathrooms:

The number of bathroom variables is statistically significant ($p < .001$). The unstandardized coefficient is 123,090.197, indicating that for every additional bathroom, the predicted home price increases by $123,090.197. The standardized coefficient is 0.538, suggesting that for every one standard deviation increase in the number of bathrooms, the home price increases by 0.538 standard deviations.

House Size (sqft):

The house size variable is also statistically significant ($p < .001$). The unstandardized coefficient is 9.265, meaning that for every additional square foot of house size, the predicted home price increases by $9.265. The standardized coefficient is 0.113, indicating that for every one standard deviation increase in house size, the home price increases by 0.113 standard deviations.

Acreage of the Lot:

The acreage of the lot variable is statistically significant ($p < .001$). The unstandardized coefficient is 22,470.156, indicating that for every additional acre of lot size, the predicted home price increases by $22,470.156. The standardized coefficient is 0.111, suggesting that for every one standard deviation increase in lot size, the home price increases by 0.111 standard deviations.

Interpretation and Insights:

The model results reveal several important insights. The significant variables, including the number of bathrooms, house size, and acreage of the lot, all demonstrate a positive relationship with home prices. This implies that larger homes with more bathrooms and larger lots tend to have higher prices in the New Jersey housing market.

Recommendations:

Based on the findings, certain recommendations can be made for stakeholders in the New Jersey real estate market:

Builders and Sellers:

Builders and sellers should consider emphasizing the number of bathrooms when marketing 3-bedroom homes, as it has the most significant impact on home prices. Additionally, highlighting the house size and acreage of the lot can also be effective in attracting potential buyers and justifying higher price points.

Buyers and Investors:

Buyers and investors interested in 3-bedroom homes in New Jersey should factor in the number of bathrooms, house size, and lot size when assessing property values. These variables play a substantial role in determining the price of homes, and considering them can help make informed decisions.

Real Estate Market Analysis:

Real estate market analysts and professionals can utilize the results of this model to gain insights into the New Jersey housing market. Understanding the importance of variables

such as bathrooms, house size, and lot size can aid in assessing market trends and property valuations.

Ethical Implications, Study Limitations, & Future Research

The data used for this study was collected from reliable sources and analyzed in an ethical manner. The use of aggregated data without personally identifiable information protects privacy and ensures ethical standards.

While there were no significant ethical limitations in this study, certain limitations should be acknowledged. The study focused on a specific region (New Jersey) and a specific type of property (3-bedroom homes), which may limit the generalizability of the findings to other regions or property types. Future research could explore the applicability of the model to different geographical areas or property categories.

Furthermore, the study's data may not capture the full complexity of factors influencing home prices. Other variables such as location, neighborhood amenities, and market conditions could also play significant roles. Including these variables in future studies could provide a more comprehensive understanding of home price determinants.

Lastly, using more recent data would enhance the model's relevance and accuracy. As the real estate market is dynamic, incorporating the latest data would capture any changes in market trends and pricing dynamics.

Despite these limitations, the findings of this research contribute to the understanding of factors influencing New Jersey 3-bedroom home prices and can serve as a valuable resource for homebuyers, real estate professionals, and policymakers.

References:

McBride, M. (2023, June 16). *What Makes Property Value Increase? 10 Factors*. Rocket Homes. Retrieved July 17, 2023, from https://www.rockethomes.com/blog/home-selling/factors-that-influence-homes-value

Data Source:

Hassan, M. (n.d.). *USA Real Estate Dataset*. Kaggle. Retrieved July 17, 2023, from https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset