From Data to Decisions: Understanding housing prices determinants

Using regression models analysis

Esiquio Iglesias Guerra *eiglesiasgue@umass.edu*

UMassAmherst

Data Analytics and Computational Social Science Program

Research Question

What is the influence of variables such as house features (# of bed, #of bathroom, house size, lot size), location (Region), median household income, density of population and crime index in determining housing prices using regression models?

This study aims to build predictive models to assess the impact of these variables on housing prices in Massachusetts, employing regression models. By analyzing the significance and magnitude of coefficients, we aim to provide insights into the factors driving housing market dynamics and inform decision-making for various stakeholders.

Hypothesis

Hypothesis #1: Larger houses is associated with higher housing prices linked to price perception, utility, prestige and status.

Sub-hypothesis:

- H1-1: The number of bedrooms positively correlates with higher housing prices.
- H1-2: The number of bathrooms is positively associated with increased housing prices.
- H1-3: Larger house sizes in square feet are positively linked to higher house prices.
- H1-4: Increased lot or property sizes are positively correlated with higher housing prices.

Hypothesis #2: The location of the house and its social and economic relationships are significantly correlated with its price.

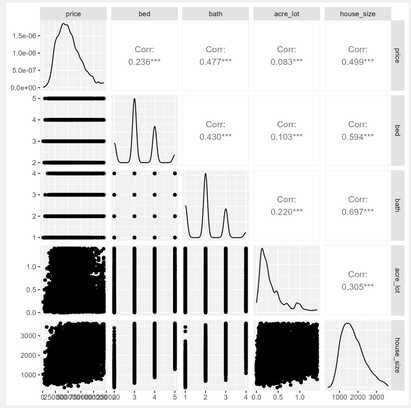
Sub-hypothesis:

- H2-1: Housing prices are higher in Western MA due to factors such as scenic landscapes, proximity to urban centers, and potentially lower crime rates compared to more urban areas.
- H2-2: Regions with higher average household incomes will have higher housing prices because individuals with higher incomes can afford more expensive housing based in status, prestige, amenities and comfort.
- H2-3: Areas with higher population densities will experience higher housing prices because they offer locations with access to amenities, employment opportunities, cultural attractions, and more.
- H3-3: Areas with higher crime rates are positively correlated with lower housing prices due to perceptions of insecurity, increased insurance premiums, and the need for costly security measures.

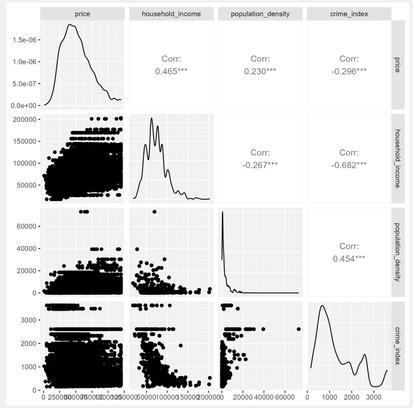
Data and Variables

- Housing features and price dataset
  - Price of houses (price), # of bedrooms (bed), # of bathrooms (bath), house size in sq feet (house_size), property/land size in acres(acre_lot)).
- Median of Household income in $ per zip code (household_income)
- Density of population per sq mile per zipcode (population_density)
- Crime index per Cities (crime_index)

Correlations



Hypothesis #1: This scatter plot shows a strong correlation between price and house size. However, it doesn't a strong linear correlation between the variables "acre lot" and price. Also, it doesn't clearly show a correlation between the variables "bed" and price, and between the variables "bath" and price.



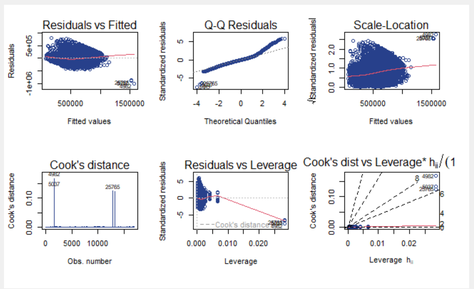
Hypothesis #2: This scatter plot shows a positive correlation between household income and price. Also, a potential trend of increasing prices with higher population density. Similarly, lower-priced houses tend to cluster in areas with lower crime indices, but the correlation is not strong.

Regression Analysis

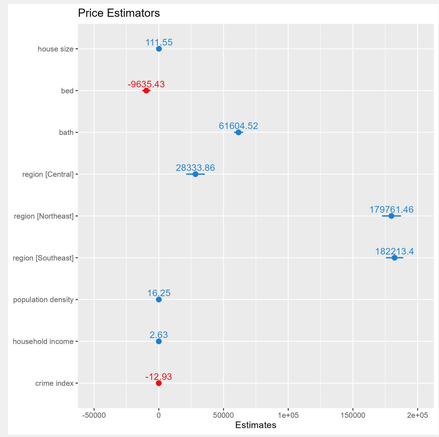
- **Hypothesis #1**
  - **Sub-hypothesis 1 :** *price = 299598 + 67793 × bed*
  - **Sub-hypothesis 2 :** *price = 239320 + 136807 × bath*
  - **Sub-hypothesis 3 :** *price = 198600 + 184.50 × sq-feet(house size)*
  - **Sub-hypothesis 4 :** *price = 497709 + 62158 × acre_lot*
- **Hypothesis #2**
  - **Sub-hypothesis 1 :** *price = 288345 + 93632×regionCentral + 349954×regionNortheast + 279733×regionSoutheast*
  - **Sub-hypothesis 2 :** *price = 207500 + 4.25 x household_income*
  - **Sub-hypothesis 3 :** *price = 481200 + 12.78 x population_density)*
  - **Sub-hypothesis 4 :** *price = 619134 - 77.59 x crime_index*
- **Final Model**

	Dependent variable:
	price
house_size	111.551*** (2.659)
bed	-9,635.435*** (1,621.481)
bath	61,604.520*** (1,801.311)
regionCentral	28,333.860*** (3,693.207)
regionNortheast	179,761.500*** (3,732.702)
regionSoutheast	182,213.400*** (3,385.241)
population_density	16.254*** (0.341)
household_income	2.626*** (0.059)
crime_index	-12.930*** (1.799)
Constant	-122,680.700*** (7,543.246)
Observations	16,916
R2	0.659
Adjusted R2	0.658
Residual Std. Error	131,867.100 (df = 16906)
F Statistic	3,622.796*** (df = 9; 16906)
Note:	*p<0.1; **p<0.05; ***p<0.01

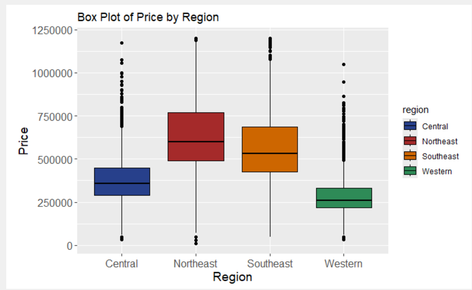
Final Model Summary: Including all the significant variables.



Final Model Diagnostics: : Points closely align around the horizontal line at 0, suggesting consistent variance across fitted values in Residuals vs Fitted Values plot. A straight line indicates normality in residuals distribution in Q-Q Residuals plot. In Cook's Distance Plot, 3 points exhibit high Cook's distance, indicating influential observations (outliers).



Plot of the results of the final regression models : This graph shows the region how the higher estimator in the price in $, also show how is the impact of the other variables.



Boxplot of Regions: This plot highlights the region's significant influence on prices. The Northeast with the highest mean housing prices, while the Western region shows the lowest.

Conclusion

Key Factors: House size, bedrooms, bathrooms, region, population density, household income, and crime index are significant predictors of house prices.

Impact on Price:

- House size: Each unit increase raises the price by $111.60.
- Bedrooms: Each additional bedroom decreases the price by $9,635.00.
- Bathrooms: Each additional bathroom increases the price by $61,600.00.
- Region: Central, Northeast, and Southeast regions have significantly higher prices compared to the Western region.
- Population Density: Price increases by $16.25 for every unit increase in population density.
- Household Income: Price increases by $2.63 for every unit increase in household income.
- Crime Index: Price decreases by $12.93 for every unit increase in the crime index.

Significance: All variables have highly significant effects on price.

Model Performance: The model explains approximately 65.85% of the variability in house prices (R-squared = 0.6585).

The analysis shows a comprehensive understanding of the factors influencing housing prices. The model indicates that several key variables significantly impact housing prices. Notably, larger house sizes, increased bathroom numbers, and higher population densities correlate with higher prices, while additional bedrooms and a higher crime index to price decreases. However, regional differences play a relevant role, with the Central, Northeast, and Southeast regions exhibiting notably higher prices compared to the reference region (Western). Household income also is as a significant predictor affecting housing prices.