Tab 1

**TSA Data Science and Analytics Documentation**

Seckinger, GA 2025 Team 209001

# **Table of Contents**

Introduction 3

Purpose 4

Methods 5

Results 6

Conclusions 14

Next Steps 15

Works Cited 16

## Introduction

Housing and real estate comprise natural resources, land, and structures within property boundaries. The market for real estate is constantly in motion, reacting to external and internal pressures such as population growth, economic policy, interest rates, and supply and demand. Although markets individually operate independently, general economic trends and policy initiatives have ripple effects on regional levels, creating alterations in affordability, pricing, and general access to housing. Furthermore, social and cultural elements such as migration patterns, lifestyle preference changes, and generational changes in homeownership values can influence housing markets. Government policies, taxation policies, and zoning regulations also have significant roles to play in determining the stability and accessibility of the market.

The COVID-19 pandemic was a major disruptor of the housing market, bringing about precipitous changes in house prices, stock in the market, and migration. Suburban and countryside locations saw a rise in demand with the adoption of working from home, while urban centers saw pricing and rental supply changes. (Felicia Di Liddo, 2023) Furthermore, higher material costs and supply chain disruptions affected new home construction, lowering inventory and driving price spikes. Volatility in mortgage rates and economic uncertainty led to periods of market heat interspersed with slow downs, thereby creating instability in buyers' confidence levels.

The impact of the pandemic raises various economic issues such as unemployment, social challenges like increased homelessness, and ethical concerns regarding housing affordability. Furthermore, higher material costs and supply chain disruptions affected new home construction, lowering inventory and driving price spikes. Volatility in mortgage rates and economic uncertainty led to periods of market heat interspersed with slow downs, thereby creating instability in buyers' confidence levels. Social and economic concerns include the struggles financially faced by a larger portion of the population as prices continue to rise, potentially leading to more widespread poverty. As urban and suburban areas struggle to flourish and rural areas face increasingly daunting living conditions, what can we expect from the real estate market and its trends?

## Purpose

The research in this paper focused on the housing market in Georgia to quantify change in housing prices and its effects on affordability and future trends. These themes can help the general population understand the challenges and dynamics of the real estate market. Due to housing affordability becoming a growing concern in many regions including Georgia, identifying patterns and trends in the costs that affect accessibility is crucial.

Poverty is a widespread issue centered on the idea of homelessness and the lack of resources to meet basic needs. According to the U.S. census bureau, 12.5% of the total U.S. population possessed income relatively below their respective poverty thresholds. (Benson, 2024) Stated by the Assistant Secretary for Planning and Evaluation (ASPE), poverty has stayed at similar levels for the past years (Office of the Assistant Secretary for Planning and Evaluation, 2023), but shows a potential increase following recent economic trends. This problem will continue to arise as the housing market increases in price and many other factors such as stagnant income and further inflation. With the greatest factor being the housing prices, it is more crucial to understand these problems and address them on a national scale.

The data was examined in 2 different features based on the information provided by the acquired datasets:

1. Price (Price, Owner-occupied units value, General Value)
2. Location (Latitude and Longitude)

The data’s analysis, conclusion, and call for next steps will nestle upon these 2 features.

## Method

We obtained our data through secondary data collection. This includes the firsthand assemblage of existing data previously collected for different or similar purposes. The datasets were acquired from the following sources: American Community Survey (ACS), American Housing Survey (AHS), Georgia Association of Realtors, ArcGIS, and Kaggle. In total, five (5) datasets in addition to other pre-existing datatables were used for the research.

Specific parameters were applied during the process of collecting data in order to align with the purpose:

1. The data must be mainly focused on Georgia
2. Data must be collected between 2019-2022 and separately in 2023 to 2024 in order to track current trends.
3. The selection criteria for datasets placed a strong emphasis on credible economic organizations, industry publications, and government sources.

After collecting the data, the data was preprocessed to fully arrive at the stage of being able to analyze the data.

* For the datasets collected from ACS:
  + Removed percentages, strings and null values using .drop([], axis=1) in Google Colab
* For the datasets collected from AHS:
  + Deleted subcategories in excel spreadsheet due to redundancy and multiple unnecessary string values.
* For the datasets collected from Kaggle:
  + Used pd.to\_datetime(df[‘time’], unit=’ms’).dt.date to change from unix to datetime values to analyze range of time
  + Dropped any null and missing, and replaced or removed string values using .drop([], axis=1)
* All original datasets were synthesized to create five different datasets reflecting the themes previously mentioned in the *Purpose* page.
* Other datatables that were retrieved have been from reliable sources, providing further insight into the analysis.

To enhance any predictions, a linear regression model was developed using structured numerical datasets (.csv format)

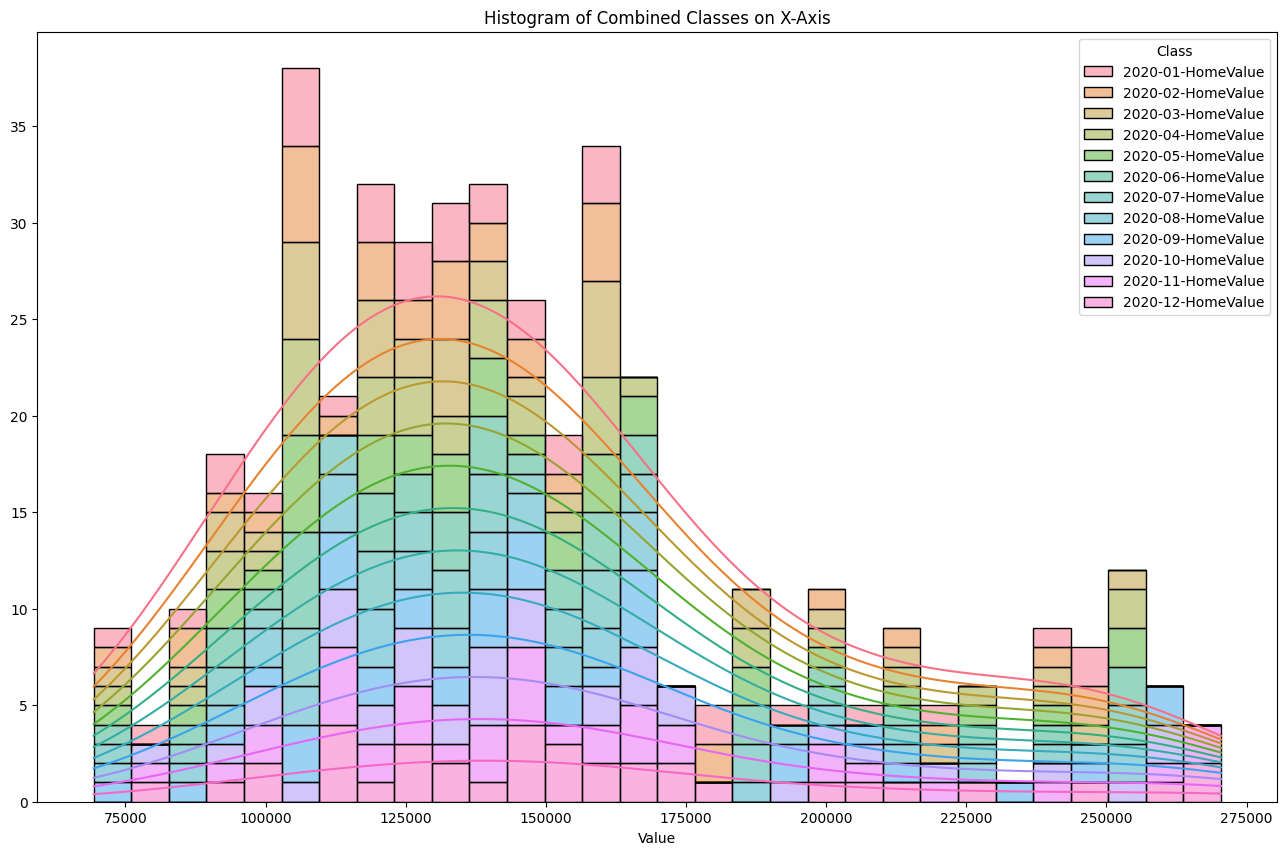
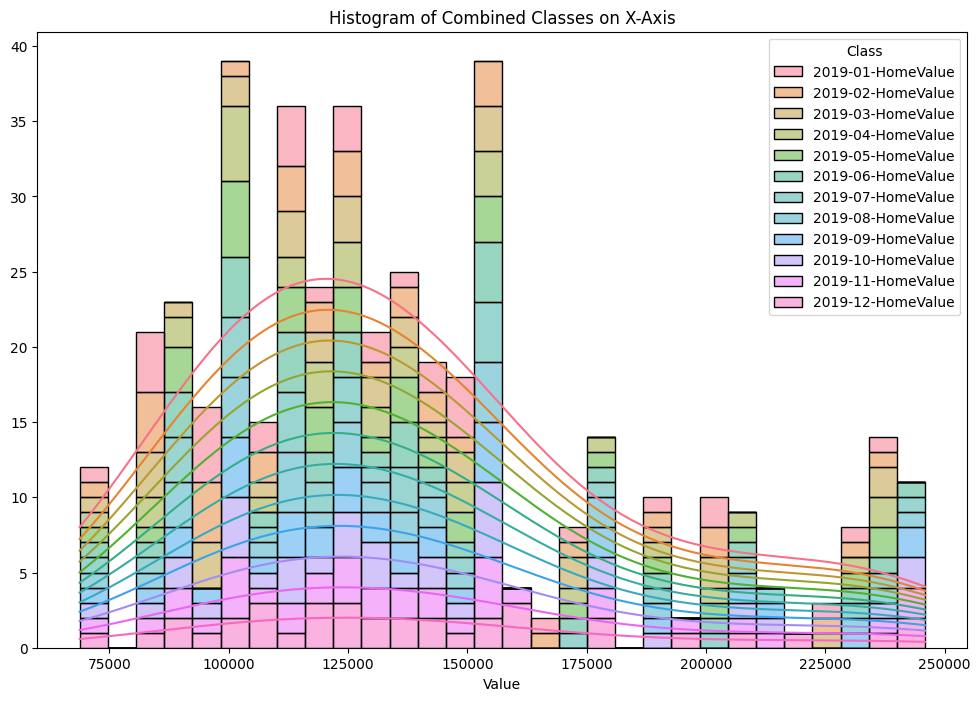
* Feature Selection:
  + Primary features included price and location in different forms
  + Categorical string values were converted to numerical representations for modeling
* Data splitting:
  + Dataset was divided into 75% training data and 25% testing data
  + The target variable (price) was separated from predictor variables to allow price prediction
* Scaling:
  + MinMaxScaler was applied to normalize numerical values and ensure consistent data scaling using scaler.fit\_transform()
* Model training:
  + Implemented a linear regression model using scikit-learn’s LinearRegression function
* Model Evaluation:
  + Assessed model performance using the respective error metrics:
    - Mean Absolute Error (MAE)
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)

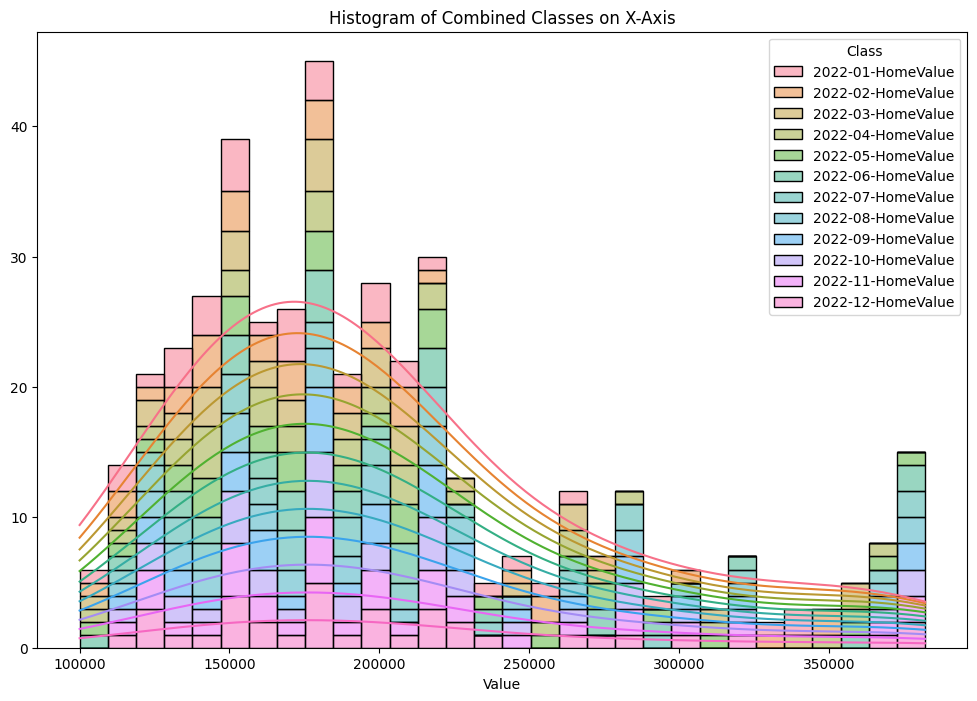
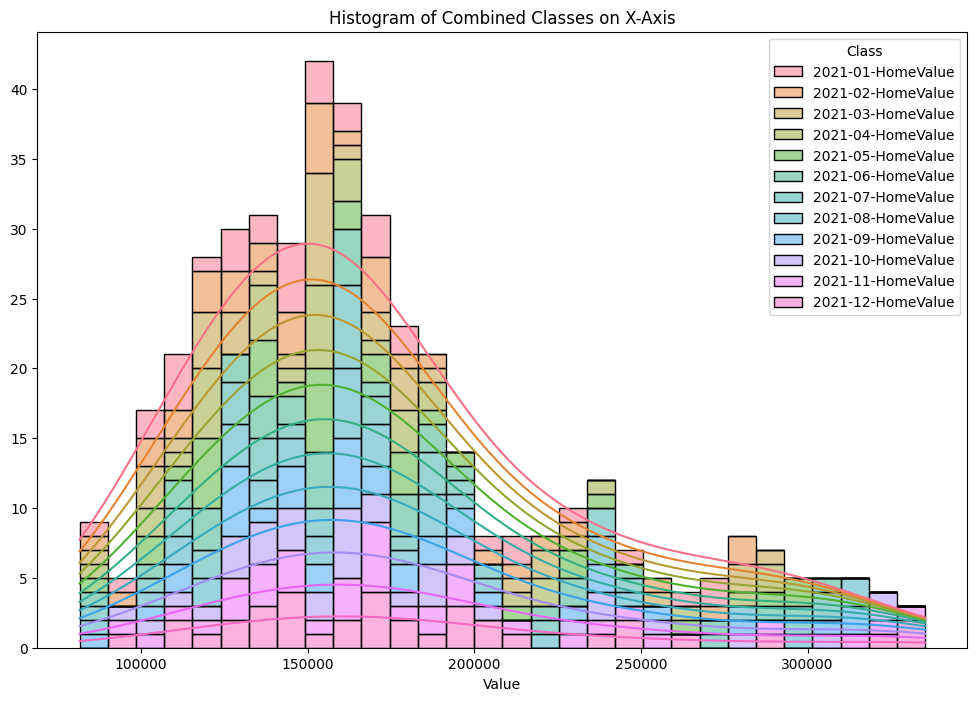
A linear regression model served primarily as a supportive tool to validate any existing trends observed in the dataset. Only previously structured .csv datasets were utilized, any excel spreadsheets were not deployed.

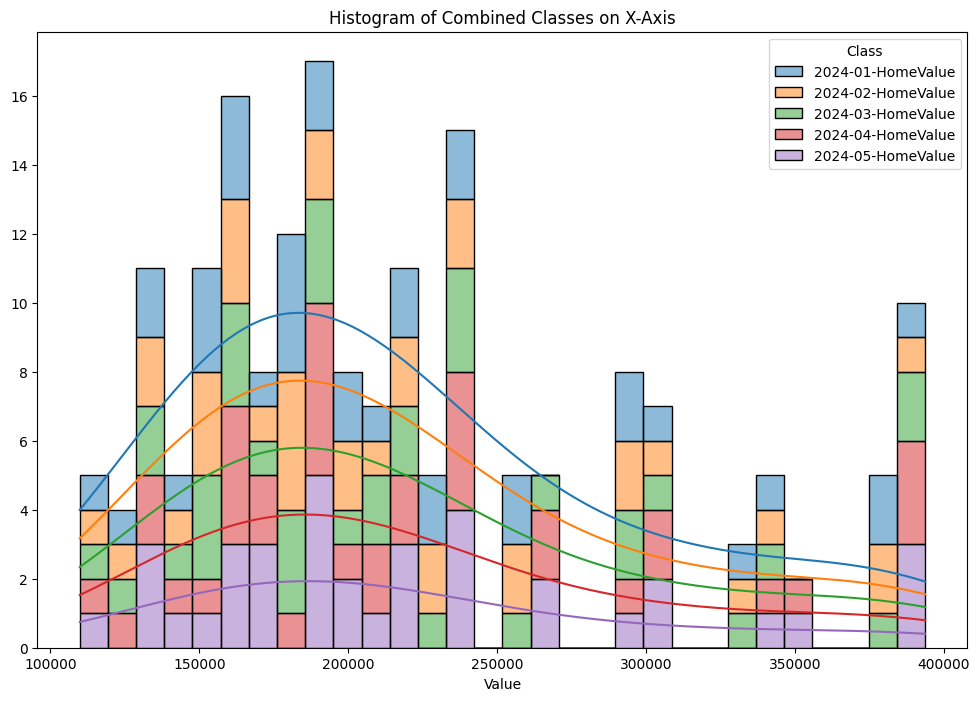
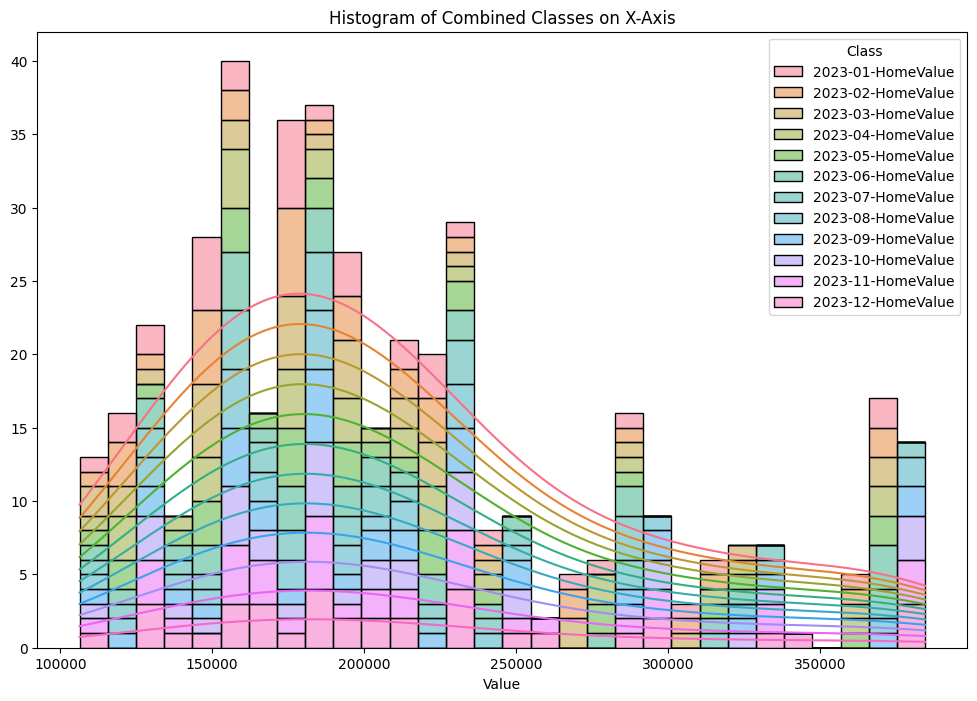
## Results

1. Price (Price, Owner-occupied units value, General Value)

The data collected for price was measured in multiple ways. Multiple sets were used, so a collective analysis was required to make sense of all prices at the time and event.







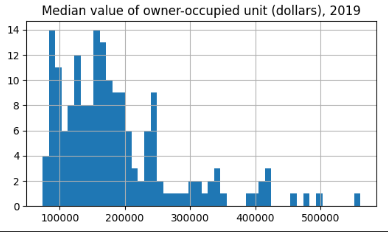
**Figure 1.** Price distribution 2019-2024 respectively from Kaggle Dataset

| Year | Mean | Median | Standard Deviation |
| --- | --- | --- | --- |
| 2019 | 138133.614454 | 128269.548988 | 44247.6328223 |
| 2020 | 148555.291803 | 138861.725507 | 47124.9037575 |
| 2021 | 172513.572398 | 159111.477794 | 55949.8459759 |
| 2022 | 202153.182641 | 183522.63645 | 69034.7691191 |
| 2023 | 211518.711851 | 193940.045304 | 72635.0346267 |
| 2024 | 219995.096555 | 200440.834042 | 76033.947465 |

**Table 1.** Mean and Standard Deviation Per Year for Value

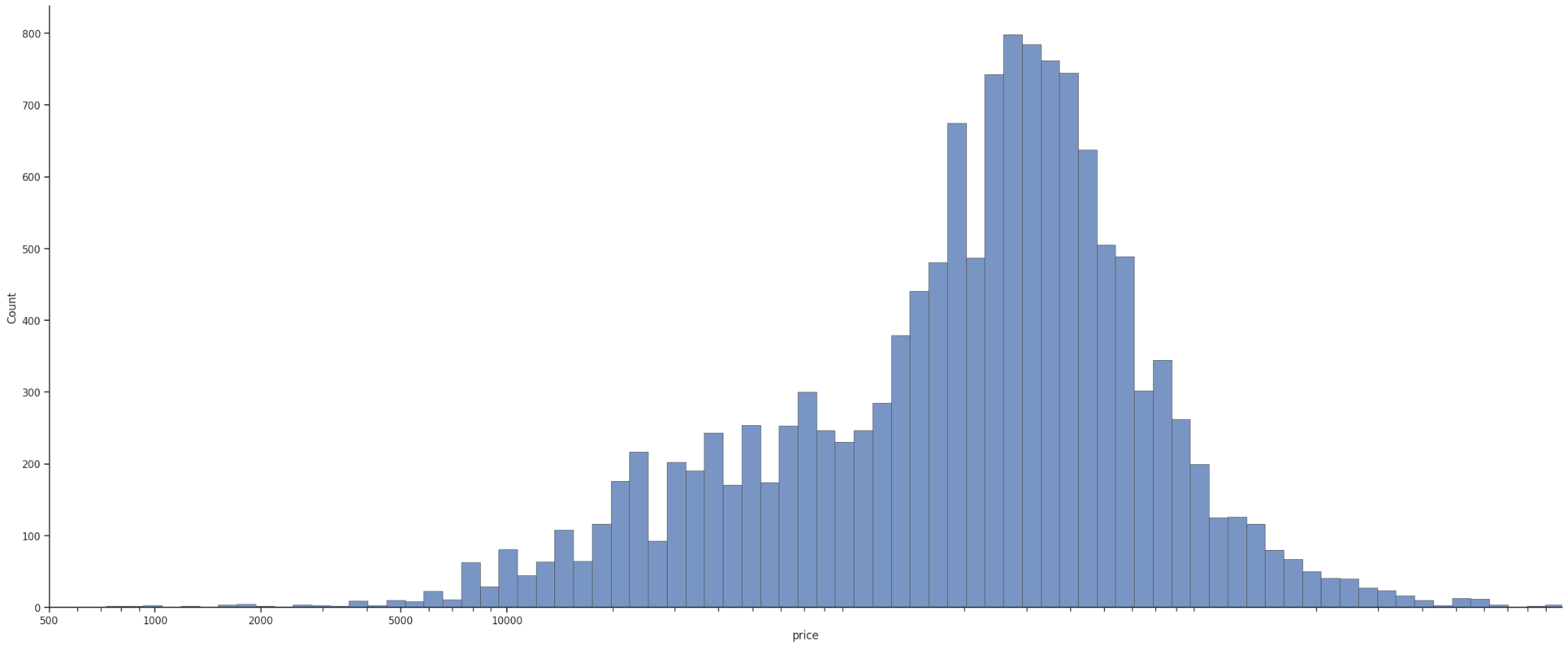
**(i)** Figure 1 and table 1 mean and median housing values consistently increase every year, indicating the increase in prices every year for the housing market

**(ii)** Median values are slightly less than the mean value, indicating that there are high house prices that skew the mean higher than the median.



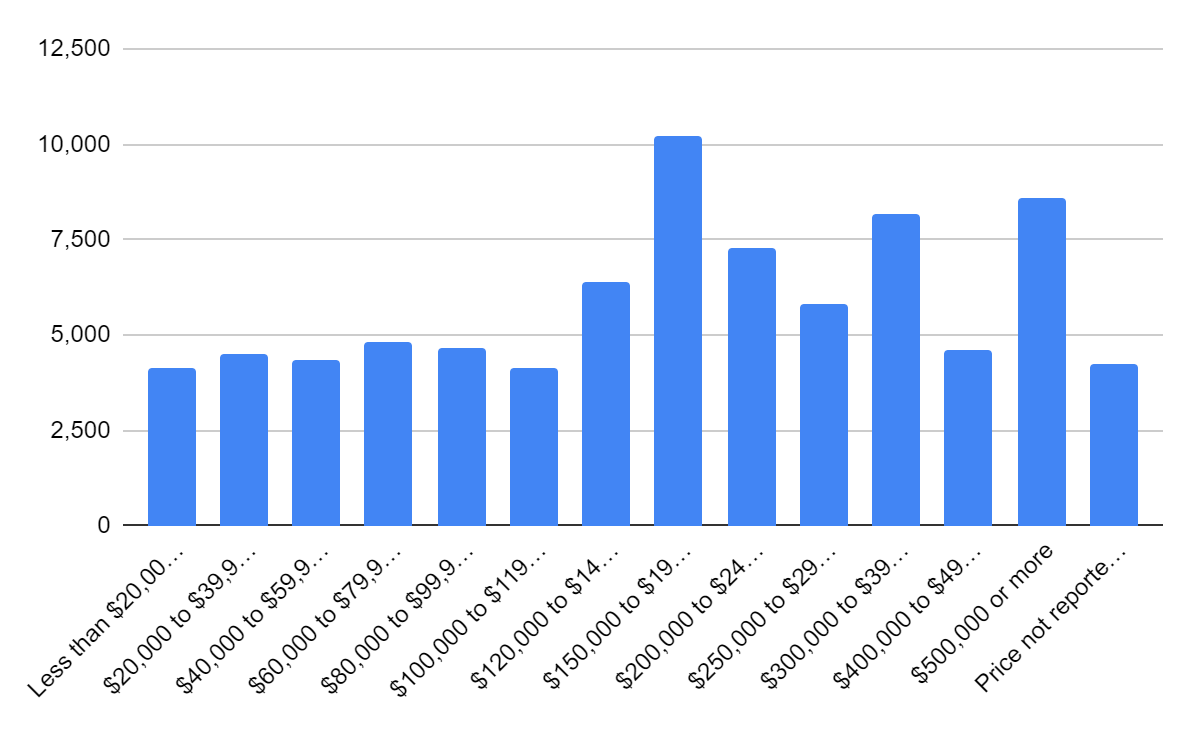
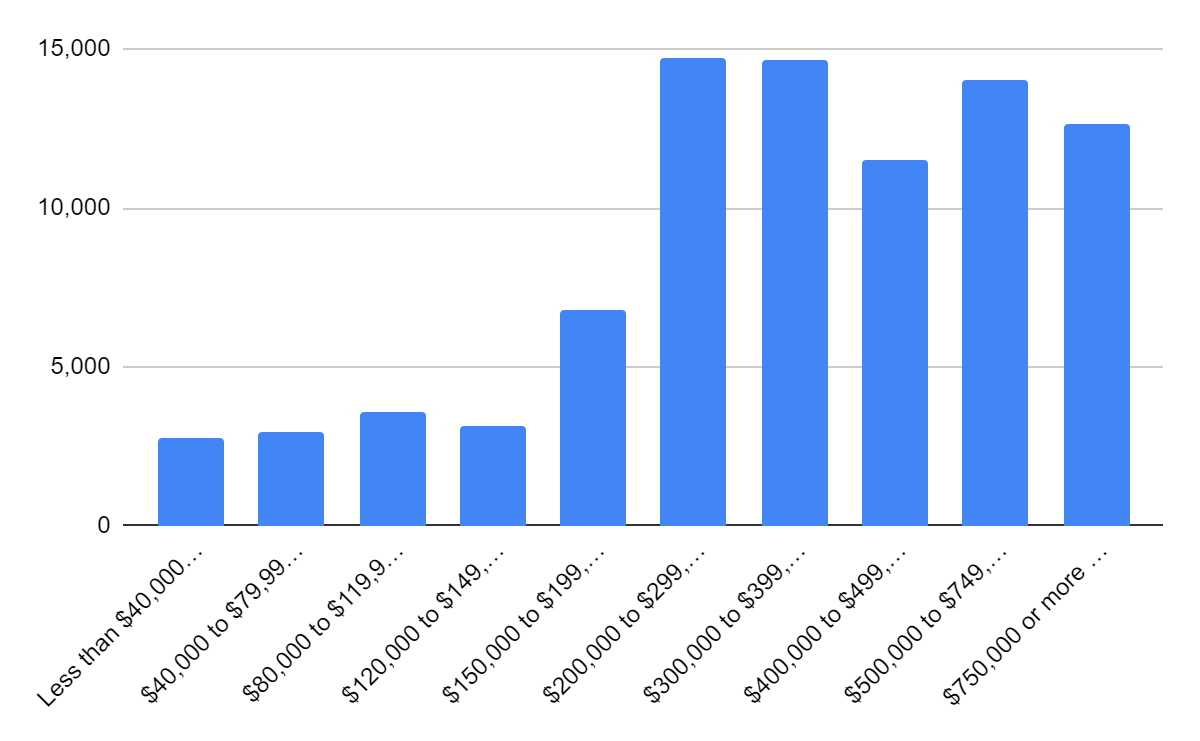
**Figure 2.** 2019 Trend of Median Values of Owner-Occupied Units in Dollar**s** from ACS

**(iii)** The median value of owner-occupied units in 2019 centered around $100,000 - $300,000 from ACS dataset



**Figure 3.** House value trend in 2019 from Kaggle Dataset

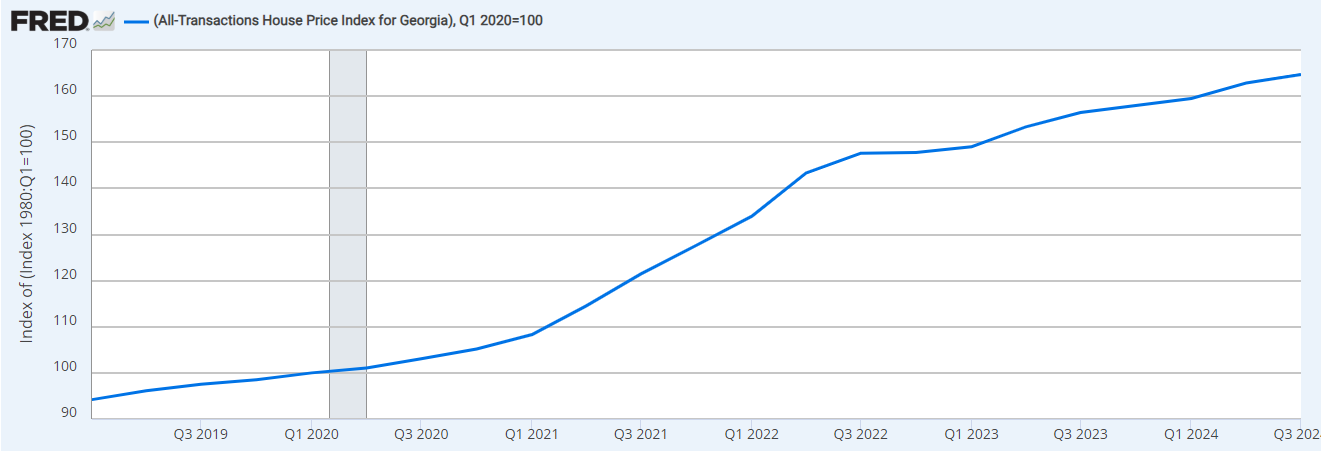
**(iv)** The median value of housing prices from Kaggle is $247,000 and is comparable to the dataset from ACS, where the median values are ranged similarly.



**Figure 3.** Histogram of 2023-2024 Housing Value and Price from AHS

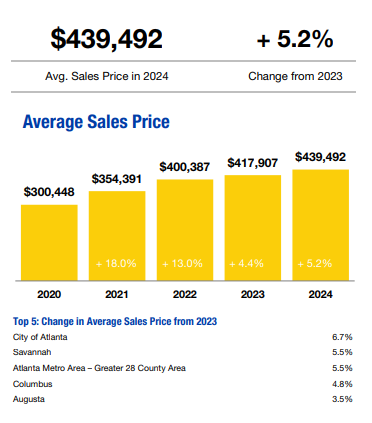
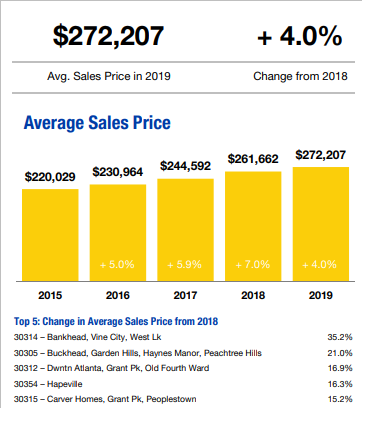
**(v)** In 2023-2024, the histogram from the United States Census Bureau indicates

more houses with higher prices, within the $300,000 - $750,000 range, roughly aligning with the trend from figure and table 1 regarding positively skewed market prices.



**Figure 4.** House Price Transaction from 2019 to 2024 from the Federal Reserve Bank of St. Louis

**(vi)** Figure 4 further supports the argument of positively skewed prices, as housing prices dramatically increase within the span of 5 years.



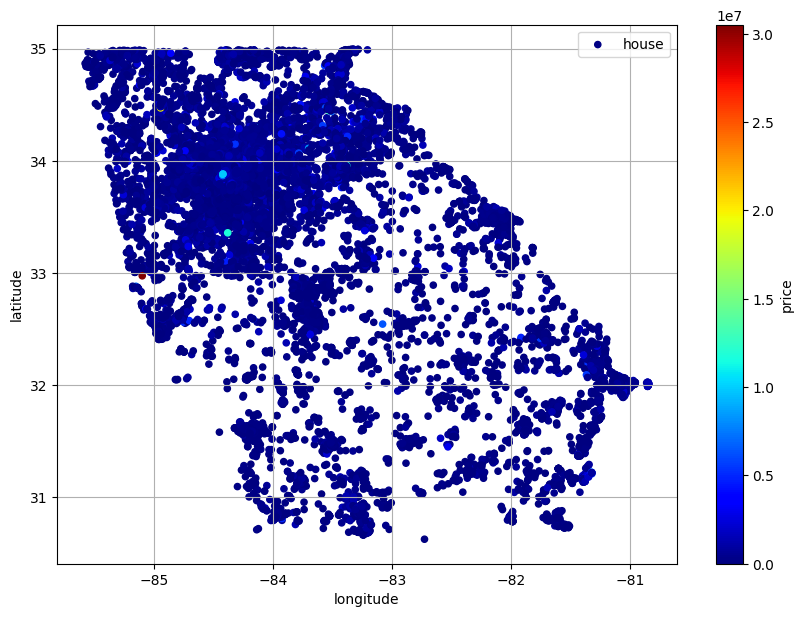
**Figure 6.** Average sales price from 2019 and 2024 respectively according to the Georgia Realtor’s Annual Report

**(vii)** Total housing transactions have increased since 2019, though only half of listed houses were sold that year.

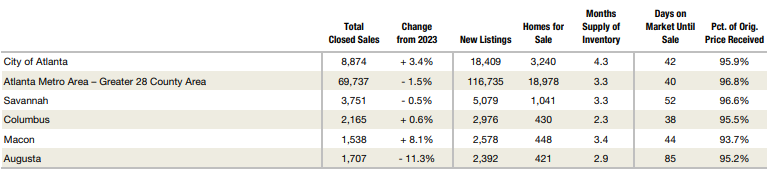
**(viii)** Average sales prices have risen post-2019, potentially due to the COVID-19 pandemic and shifting supply-demand dynamics

This data suggests affordability constraints have led to price stabilization at lower tiers, while high-end homes continue to appreciate. Figure 6, roughly aligning once again with the trends from Figure 1, indicates gradually higher sales over the years. There are a few mathematical inconsistencies throughout the collected data, where some values do not align with values given by other datasets and contain data that may be skewed, which could be from artificially generated data or other factors that were not considered. However, they all still represent similar trends that come to similar conclusions.

1. Location (Latitude and Longitude)

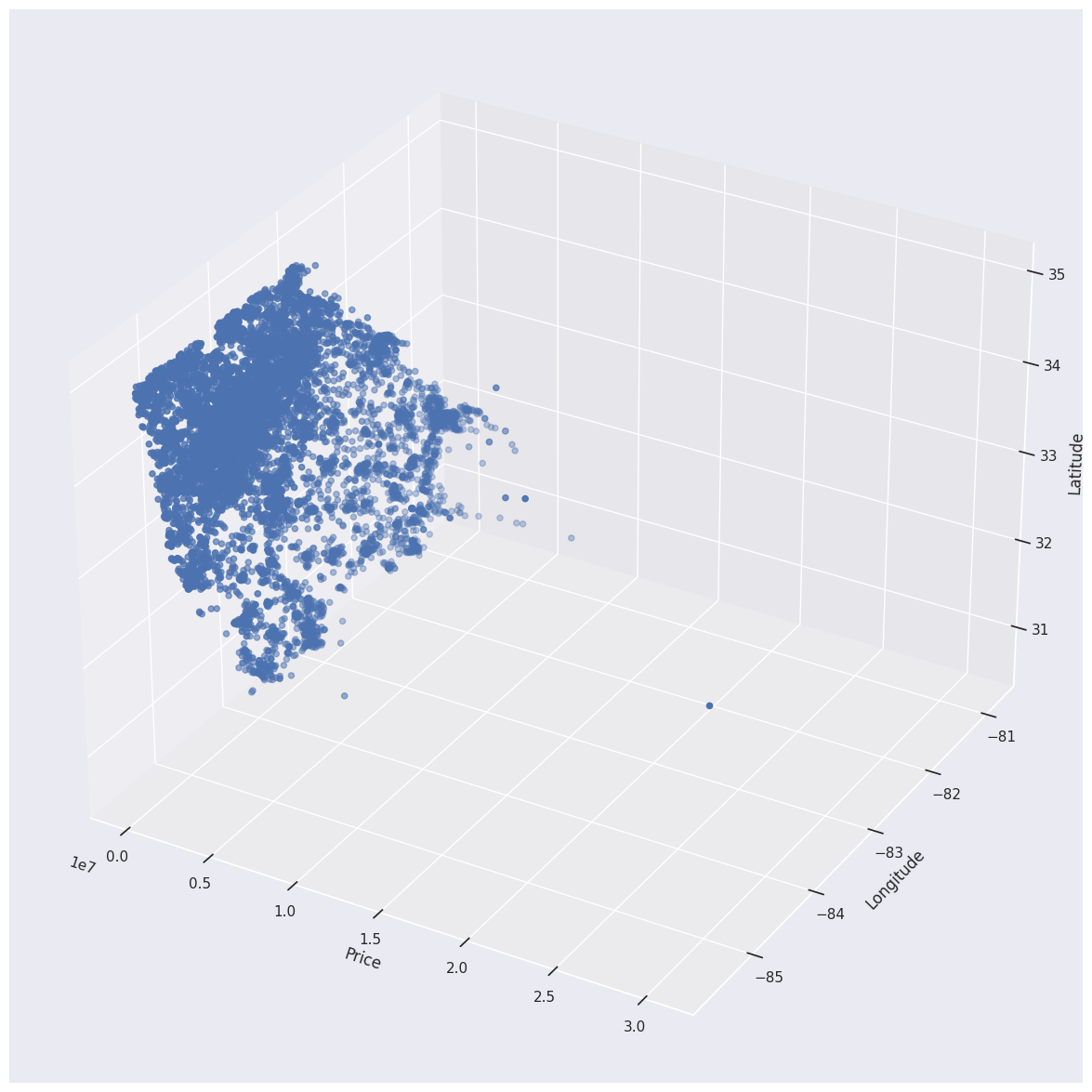


**Figure 7.** Scatter plot graph of occupational density in Georgia 2019 from Kaggle



**Figure 8.** Area overview 2024 - Georgia Association of Realtors

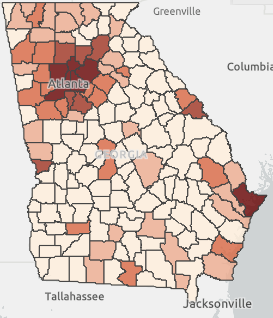
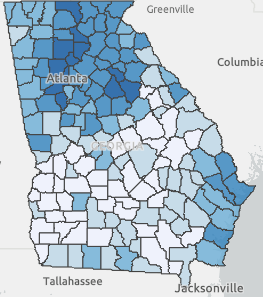
**(i)** Figure 7 displays the occupational density of houses in 2019, displaying population density being mainly concentrated around major cities.



**Figure 9.** Price to Latitude and Longitude values

**(ii)** Figure 9 displays the higher price generation around the major cities, and figure 8 further supports this conclusion as changes are higher from the previous year.

[A] [B]

****

**Figure 10.** House to population density in 2022 to 2024 from ArcGIS

(iii) Comparative to figure 10, housing to population density has stayed relatively the same across 2022 to 2024 and consistent to trends shown from figure 7 indicating a consistent rise in price and value in urban areas.  
This presents a market centered around urban centers and more populated city areas, which continually show a rise in price according to figure 8 by a certain percentage each year, and will continue to do so. Inconsistencies within the graphs may be present as these datasets pull from different backgrounds: demographics, numerical representation, quantity of collected data.

1. Linear Regression Machine Learning Model

A process of linear regression was used for one of the datasets from Kaggle’s Georgia Real Estate from 2019 to perform a predictive analysis on market value price.



**Figure 11.** Correlation Matrix of Scaled Values

A correlation matrix of the scaled values can reveal trends and connection to each other.

**(i)** Price and longitude shows a negative and weak correlation indicating that price and longitude have semi-relatable values, but still weak associations.

**(ii)** Price and latitude shows a positive and weak correlation indicating that price and latitude also have semi-relatable values, but still weak association.

**(iii)** Longitude and latitude correlations are mostly irrelevant because they are location values and do not concern price.

**Table 2.** OLS Regression Results

| Dep. Variable: | price | R-squared: | 0.011 |
| --- | --- | --- | --- |
| Model: | OLS | Adj, R-squared: | 0.011 |
| Method: | Least Squares | F-statistic: | 57.34 |
| No. Observations: | 10353 | Prob (F-statistic): | 1.71e-25 |
| Df Residuals: | 10350 | Log-Likelihood: | 24676. |
| Df Model: | 2 | AIC: | -4.935e+04 |
| Covariance Type: | nonrobust | BIC: | -4.932e+04 |

|  | coef | Std error | t | P > |t| | [0.025 | 0.975] |
| --- | --- | --- | --- | --- | --- | --- |
| const | 0.0066 | 0.001 | 6.520 | 0.000 | 0.005 | 0.009 |
| longitude | -0.00007 | 0.001 | -0.597 | 0.551 | -0.003 | 0.002 |
| latitude | 0.0093 | 0.001 | 8.672 | 0.000 | 0.007 | 0.011 |

| Omnibus: | 21090.789 | Durbin-Watson: | 2.034 |
| --- | --- | --- | --- |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 117599319.176 |
| Skew: | 16.719 | Prob(JB): | 0.00 |
| Kurtosis: | 524.054 | Cond. No. | 9.73 |

**(iv)** Table 1 summarizes the model fit where the low R-squared (0.011) value tells us that the model explains very little of the variance in house prices.

**(v)** Additionally, the significant predictor seems to be the latitude value, and not the longitude value.

**(vi)** A residual1 analysis of high skew and kurtosis values with significant Omnibus and Jarque-Bera tests indicate that the residuals of the data are not normally distributed and potentially affect the reliability of inferences made throughout the data.

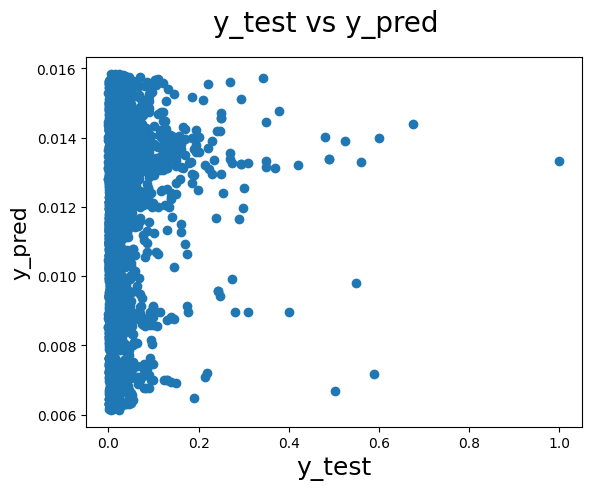
**Table 3.** Error Values from Model

|  | Mean Absolute Error (MAE) | Mean Squared Error (MSE) | Root Mean Squared Error (RMSE) |
| --- | --- | --- | --- |
| Respective Metrics | 0.03 | 0.00 | 0.06 |

**(vii)** The MAE, MSE, and RMSE values indicate a low level error value, indicating that the model performed well in its accuracy.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Residual - Difference between predicted and actual values



**Figure 12.** Model prediction results

Overall, the low R-squared values from table 2 indicate that the model does not explain much variation in house prices, and while it is statistically significant, it is not effective at predicting prices and thus needs additional features to improve performance. The latitude value shows a statistically significant impact on house price, while longitude does not. This suggests that the geographical location of latitude plays a role in determining house prices in the dataset, but the longitude value has little to no influence. The high skew and kurtosis indicate that the residuals are not normally distributed, and can be observed in figure 12. This will be problematic because many statistical models, including OLS, assume that the residuals are normally distributed. The high skew suggests that the distribution is asymmetric and high kurtosis indicates presence of outliers, causing issues in the final model.

Full summary and details of the table are listed below:

The general model consists of:

* **Dependent Variable: price** - Variable being predicted
* **R-squared: 0.011** - Indicates 1.1% of the variability in the house prices is explained by the predictors in the model (longitude and latitude) and a score of this degree hints that the model doesn’t explain much of the variation in price.
* **Adj. R-squared: 0.011** - Adjusts R-squared for the number of predictors relative to the number of data points, and confirms that the predictors don't explain much additional variance.
* **Model: OLS (Ordinary Least Squares)** - Type of regression analysis performed, minimizes sum of squared errors between observed and predicted values.
* **Method: Least Squares** - Indicates that the estimation method used is the least squares approach
* **No. Observations: 10353** - Total number of data points (houses) used in the regression.
* **Df Model: 2** - Number of predictors (longitude and latitude) in the model
* **Df Residuals: 10350** - Degrees of freedom left after estimating the model, calculated as the total number of observations minus the number of estimated parameters
* **F-statistic: 57.34 and Prob(F-statistic): 1.71e-25** - Tests whether at least one of the predictors is statistically significant. A very low value indicates that the overall model is statistically significant despite the low R-squared value.
* **Log-Likelihood: 24676** - A measure of how likely it is that the model would produce the observed data.
* **AIC (Akaike Information Criterion)**: -49350 and BIC (Bayesian Information Criterion): -49320 - Criteria for model selection that penalize model complexity. Lower values suggest a better fit.
* **Covariance Type: nonrobust** - Indicates assumption made about the error variance. Nonrobust means standard errors are calculated under the assumption that the errors are homoscedastic (constant variance).

The coefficients table consists of:

* **Const (intercept): 0.0066** - Predicted value of the dependent variable when all independent variables are zero
* **Longitude: -0.00007** - Coefficient for longitude, where a one-unit increase is associated with a decrease in price of 0.00007 units, holding all else constant. However, the p-value notes that this relationship is not statistically significant
* **Latitude: 0.0093** - Coefficient for latitude, where a one-unit increase is associated with an increase in price of 0.0093 units, and this relationship is statistically significant.
* **Std error** - The standard error of the coefficient estimate, reflecting the variability of that estimate
* **T** - the t-statistic determines whether the coefficient is significantly different from zero
* **P > |t|** - The p-value for the t-test, where a small value (typically < 0.05) indicates that the coefficient is statistically significant
* **[0.025, 0.975]** - The 95% confidence interval for the coefficient, suggesting a 95% confidence that the true coefficient lies within this range.

The additional statistics include:

* **Omnibus: 21090.789 and Prob(Omnibus): 0.000** - Test for skewness and kurtosis of residuals. A significant result (low p-value) indicates that residuals deviate from a normal distribution
* **Durbin-Watson: 2.034** - Tests for autocorrelation in the residuals. A value close to 2 suggests that there is no autocorrelation.
* **Jarque-Bera (JB): 117599319.176 and Prob (JB): 0.00** - Another test for normality of residuals, where a high JB value with a p-value of 0.00 suggests that residuals are not normally distributed and is further indicated by the high skew and kurtosis
* **Skew: 16.719** - Measures the asymmetry of the residual distribution, where a high positive skew indicates a long tail on the right side.
* **Kurtosis: 524.054** - Measures the “tailedness” of the residual distribution, and a very high kurtosis suggests that the distribution has heavy tails or outliers.
* **Cond. No.: 9.73** - Condition number, potentially indicating multicollinearity issues among predictors, where a low condition number suggests no serious multicollinearity.

## Conclusions

1. Price

An analysis of housing prices from 2019 to 2024 demonstrates an upward trend with both mean and median values increasing yearly (Table 1, Figure 1). The consistent rise in prices suggests a market that favors appreciation, particularly in high-priced housing, as evidenced by the positively skewed price distributions (Figures 3, 4, and 6). The surge in housing prices is likely influenced by events such as the COVID-19 pandemic and changing supply-demand dynamics. While some inconsistencies exist across the datasets, the overall trend indicates that home values will likely continue to rise, with affordability constraints impacting lower-income buyers while more costly properties will maintain strong appreciation.

1. Location

Housing prices are closely tied to location, with data showing that price increases are more noticeable in densely populated urban centers (Figures 7-10). Major civilized areas have consistently higher housing costs, supporting the correlation between location and property value. Population and occupational density data from 2019 to 2024 suggest that the demand for urban housing remains strong, leading to a sustained price growth near city centers. While suburban and rural areas experience price increases as well, they are not as drastic compared to urban hubs. The data further suggests that as urban expansion continues, housing prices will also likely follow the same upward trajectory, with location remaining a key factor in price determination presumably from multiple factors such as events (COVID-19, Super Bowl, etc.) and biases regarding living conditions, neighborhood, demographics, and many more.

1. Linear Regression Machine Learning Model

The machine learning model revealed inconsistencies throughout the data used in the Kaggle dataset, as well as inconsistencies within the features used. While the analysis overall was performed throughout multiple datasets and considered multiple viewpoints, the model itself used little features that included insufficient amounts of features and affected the model. However, using more features as well as additional data will provide a more accurate and complex model that would be accurate enough to predict prices in the near future. Additionally, although the model failed to accurately predict using the given features, it still revealed aspects of price and location correlations and showed a relationship that value and price is more pivoted towards the east than the west, according to the latitude to price correlation.

## Next Steps

Because the primary purpose of this research was to observe past trends of the housing market, further analysis must be done over a larger dataset, in context to more features regarding biases due to the effects of personal opinion and word of mouth in the housing economy. Additionally, steps should be taken to see if any themes of this research were by coincidence or by correlation.

Most importantly, more information should be taken in order to minimize factors such as city demographics and media influence. Furthermore, there should be a wider comparison to the U.S. housing market because of the influence the global market and economy may have on the individual financial state of housing. While it may not seem to have a great impact on the overall analysis, it can still have a feature that may influence the market as a whole that was not considered in the research previously. Amidst the pandemic, there have been other factors, such as celebrities, championships and competitions, and more factors that may skew the overall results away from a more detailed and functional examination. There are also further ways to improve the investigation, such as

Further Data Collection and Preprocessing:

* Add more datasets from 2025 onward to monitor post-pandemic trends.
* Revisit cleaning protocols to ensure even greater data accuracy.

Enhanced Predictive Modeling:

* Develop more sophisticated machine learning models using the cleaned datasets to forecast future trends.
* Use further layered feature learning and scaling techniques to refine predictions.

Detailed Regional Analysis:

* Conduct localized studies in identified hotspots (e.g., northwest Georgia) to understand the underlying factors.
* Engage with the local community to verify and supplement data findings.

Policy Recommendations and Outreach:

* Prepare detailed reports aimed at policymakers that highlight areas needing affordable housing goals/initiatives.
* Organize stakeholder meetings and public forums to discuss implications and strategies based on the analysis.

## References

“18 Million Americans Are House Poor, New Study Shows - CBS News.” *Www.cbsnews.com*, 18 June 2024, www.cbsnews.com/news/home-prices-sales-house-poor-mortgage/.

“ArcGIS Insights.” *Arcgis.com*, 2025, insights.arcgis.com/#/view/6068d9f84bd54a89b911f0c65dfd3108. Accessed 13 Feb. 2025.

“Atlanta, Georgia Housing Market Report January 2025 - Rocket Homes.” *Rocket.com*, 2025, rocket.com/homes/market-reports/ga/atlanta. Accessed 13 Feb. 2025.

Benson, Craig. *Poverty in States and Metropolitan Areas: 2023*. 2024.

Bureau, US Census. “Data.” *Census.gov*, www.census.gov/programs-surveys/ahs/data.html.

---. “Housing Data Tools.” *Census.gov*, 22 Nov. 2021, www.census.gov/topics/housing/data/tools.html. Accessed 13 Feb. 2025.

Centers for Disease Control and Prevention. “COVID-19 Timeline.” *Centers for Disease Control and Prevention*, CDC, 2023, www.cdc.gov/museum/timeline/covid19.html.

“Data Tables | GA Data.” *Georgiadata.org*, georgiadata.org/data/data-tables.

Felicia Di Liddo, et al. “The Impacts of COVID-19 on Real Estate Market Dynamics: A Systematic Literature Review of Emerging Trends.” *Buildings*, vol. 13, no. 9, 14 Sept. 2023, pp. 2334–2334, https://doi.org/10.3390/buildings13092334.

*FOR RESIDENTIAL REAL ESTATE ACTIVITY in the STATE of GEORGIA Annual Report on the Georgia Housing Market*. 2019.

“GEORGIA: Comprehensive Housing Market Analysis Reports | HUD USER.” *Huduser.gov*, 2025, www.huduser.gov/portal/chma/ga.html?utm\_source=. Accessed 13 Feb. 2025.

“Housing Values (by Georgia House) 2019.” *Atlantaregional.com*, 2019, opendata.atlantaregional.com/datasets/housing-values-by-georgia-house-2019/explore?location=34.828001%2C-65.734075%2C5.11. Accessed 13 Feb. 2025.

Miner, Brandie. “RECAP: Georgia Housing Market in 2024» Georgia Association of REALTORS®.” *Georgia Association of REALTORS®*, 31 Jan. 2025, garealtor.com/recap-georgia-housing-market-in-2024/. Accessed 13 Feb. 2025.

Office of the Assistant Secretary for Planning and Evaluation. “Poverty Guidelines.” *ASPE*, U.S. Department of Health and Human Services, 19 Jan. 2023, aspe.hhs.gov/topics/poverty-economic-mobility/poverty-guidelines.

“Real Estate Georgia.” *Www.kaggle.com*, www.kaggle.com/datasets/yellowj4acket/real-estate-georgia.

US Census Bureau. “Poverty.” *Census.gov*, 12 Sept. 2018, www.census.gov/topics/income-poverty/poverty.html.

“USA Real Estate Dataset.” *Www.kaggle.com*, www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset.

“Annual Report on the Georgia Housing Market”

Tab 2