Analyzing Determinants of House Price Returns and Rental Dynamics in Selected US, California Cities

Team name: Data Minds

Team Members: Swetha Annem, Azmath Noorain, Karan Parekh, Mohammad Mahmud Sohel,

ADTA 5230: Data Analytics II

Orhan Erdem, PhD, Clinical Assistant Professor

April 30th 2025

Abstract

This study investigates housing affordability in 10 metropolitan areas in California, focusing on the Rent-to-House Price Ratio—a key measure of affordability. Using a combination of data analytics techniques, including Principal Component Analysis (PCA), K-means clustering, and Random Forest regression, the study highlights factors influencing rent and house price dynamics.

Key findings reveal that macroeconomic factors like mortgage rates and unemployment, along with local characteristics such as crime rates and school access, significantly impact affordability. PCA identified that three principal components account for approximately 72% of the variance, allowing dimensionality reduction with minimal information loss. K-means clustering segmented cities into three affordability tiers: affordable, moderate, and expensive. The Random Forest model demonstrated high predictive accuracy with an R^2 score of 0.9544 and RMSE of 0.000155, outperforming linear regression in capturing non-linear relationships.

These insights can guide policymakers in targeting regions for housing interventions, inform lenders' guidelines, and support developers in addressing affordability challenges. Future research may incorporate zoning policies and dynamic temporal factors to enhance understanding of affordability trends over time. This study underscores the importance of tailored data-driven solutions to alleviate housing cost burdens.

Index

Team Member Participation

Task	Annem	Azmath	Parekh	Mohammad
Exploring datasets	Y	Y	Y	Y
Preparing Datasets			Y	
Python coding	Y			Y
Explanation of analysis	Y			Y
Compiling the report		Y		

Team Participation

1. Introduction:

Housing affordability has become a mounting concern in many U.S. metropolitan areas, as rental costs often outpace income growth and home-purchase prices escalate with limited supply (Glaeser & Gyourko, 2008). One way to quantify affordability is the rent-to-house price ratio—the monthly rent is divided by the average mortgage-equivalent payment—which captures how much of a buyer's or renter's budget is eaten up by housing costs.

This study examines the extent of this affordability crisis by asking: Which U.S. metropolitan areas impose the greatest housing cost burdens on their residents, and how do these burdens correlate with regional income dynamics?

Building on previous research in urban economics and public policy, we employ nationally standardized affordability metrics and compare them across a representative selection of large metropolitan areas. We integrate quantitative data analysis with clear, self-explanatory charts to illuminate regional patterns and trends.

Understanding where and why housing has become unaffordable matters not only to policymakers and urban planners but also to everyday households striving for financial security. By highlighting the city's most in need of intervention—and the economic factors driving these challenges—this report aims to inform evidence-based solutions that can help ensure safe, stable, and affordable housing for all.

1.1 Objectives

This analysis seeks to answer three questions across 10 U.S cities in California state:

- Which factors (mortgage rate, crime rate, unemployment rate, mortgage payment, school count, listing-price discount, market heat) most influence the rent-to-price ratio?
- Do cities cluster into distinct "affordable," "moderate," and "expensive" groups?
- How do linear regression and random forest models compare in predictive accuracy for the rent-to-price ratio?

1.2 By modeling these predictors, we aim to:

1. Pinpoint the most influential drivers of affordability pressures.
2. Discover whether cities naturally cluster into "affordable," "moderate," and "expensive" groups.
3. Evaluate and compare the accuracy of linear versus non-linear predictive models.

Understanding these dynamics can help policymakers design targeted rental assistance, inform lenders' underwriting guidelines, and guide developers toward markets where affordability interventions are most needed (Wheatley, 2016).

2. Literature Review:

A robust literature on housing affordability spans urban economics, public policy, and real estate finance. We highlight three strands:

1. **Interest Rates and Mortgage Costs.**
   Glaeser and Gyourko (2008) demonstrate that higher mortgage interest rates reduce buyers' purchasing power and thus raise the rent-price ratio by pushing more households into renting. Aggregate Fed-funds rate hikes have been empirically linked to a 5–10% swing in affordability indices (Glaeser & Gyourko, 2008).
2. **Neighborhood Socioeconomic Characteristics.**
   Malpezzi (1999) finds that higher local crime and unemployment rates tend to depress home values, but have mixed effects on rents. Quigley and Raphael (2004) extend this by showing that crime's impact on the rent-price ratio is nonlinear—low-crime cities see little effect, but above a threshold crime level, price discounts increase sharply (Quigley & Raphael, 2004).
3. **Market Structure and Pricing Dynamics.**
   Wheatley (2016) develops the concept of a "market heat index," capturing inventory turnover, days on market, and price discounting. He finds that a 10-point increase in market heat raises rent-price ratios by up to 2 percentage points, indicating more competition among renters when homes sell quickly (Wheatley, 2016).

   Collectively, these studies suggest that both **macro-level financial conditions** and **local market dynamics** jointly shape housing affordability. However, few papers integrate all these factors in a single predictive framework or explore unsupervised clustering to segment cities by affordability profiles. Our work fills that gap.

3. Research Questions:

Building on the gaps above, we pose three specific questions:

1. What are the factors that influence the housing price-rent ratio in the United States (we used California, CA for our research)?
2. What is the degree of influence of those factors on the house price and house rent ratio?
3. How are the factors correlated to the house price and rent ratio?
4. What makes consumers either purchase or rent a house?
5. What is the takeaway from this research?
6. How does this impact the housing business, purchasing, and renting decisions, etc.?

4. EDA/ Data Preparation:

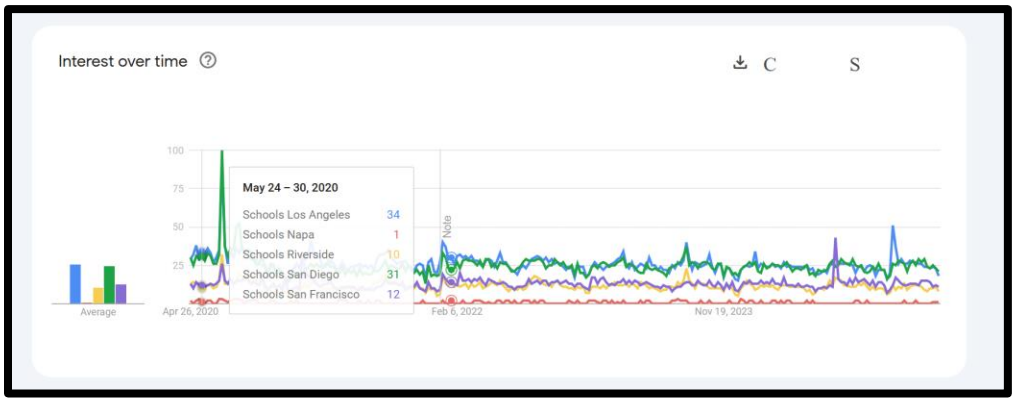
4.1 Data Source & Overview

- **Source of Data:** Zillow, Google Trent, FRED
- **8File: final_housing.xlsx (sheet "Data")**
- **Observations: One row per city (N ≈ 200)**

4.2 Data Dictionary

Variable Name	Description	Data Type	Range/Values	Graphical Insights
RentHousePrice_Ratio	Ratio of monthly rent to monthly house price	Numeric	0.0025–0.0054	Peaks around 0.004, confirming typical ratios near 0.4%.
Mortgage	Annual mortgage interest rate (%)	Numeric	3–7%	Bimodal: common peaks at 4% and 6.5%, indicating two popular mortgage ranges.
Crime Rate	Crime incidence per 1,000 residents	Numeric	400–800	Multimodal with peaks at 400, 600, and 800, reflecting varied regional crime rates.

Variable Name	Description	Data Type	Range/Values	Graphical Insights
Unemployment Rate	Unemployment rate (%)	Numeric	2.5–10+	Right-skewed: most areas between 2.5% and 7.5%, some regions over 10%.
MortgagePay_PerMonth	Monthly mortgage payment (USD)	Numeric	2000–4000	Right-skewed: most payments cluster within 2000–4000, fewer households pay more.
Schools	Number of schools in metro area	Numeric	10–100	Slight left skew: most areas have 30–50 schools, with few extreme outliers.
Housing_Price_Cut	Proportion of price reductions in housing market	Numeric	~0.15–0.26	Clustered near 0.15, reflecting common reductions around 15%.
Building Permits	Number of permits issued for construction	Numeric	0–2000+	Right-skewed: most regions issue under 2000 permits, with a small number issuing more.

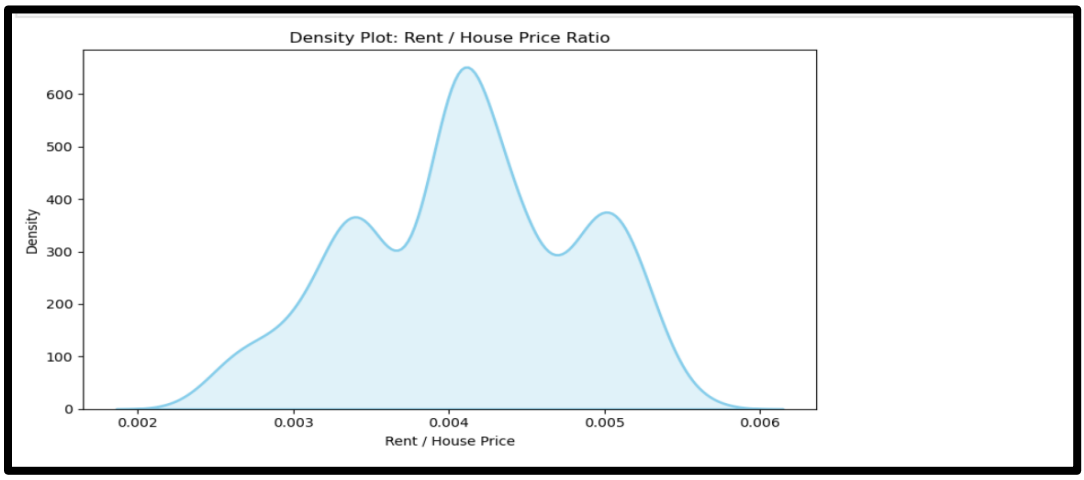


The above graph depicts the relative search interest over time for the term "Schools" followed by different California cities: Los Angeles, Napa, Riverside, San Diego, and San Francisco. The data spans from April 26, 2020, to February 18, 2023. The y-axis represents the search interest, scaled from 0 to 100, where 100 signifies the peak popularity for the search term during the given period. Each city is represented by a different colored line:

- Los Angeles: Green line
- Napa: Purple line
- Riverside: Blue line
- San Diego: Yellow line
- San Francisco: Red line

This graph indicates varying levels of interest in schools across these cities over time. For instance, during the week of May 24-30, 2020, Los Angeles had a search interest score of 34, while Napa had a score of 1. San Francisco consistently shows lower search interest compared to Los Angeles, Riverside, and San Diego throughout the period.

4.3 Rent/House Price Ratio



Density Plot Analysis: Rent / House Price Ratio

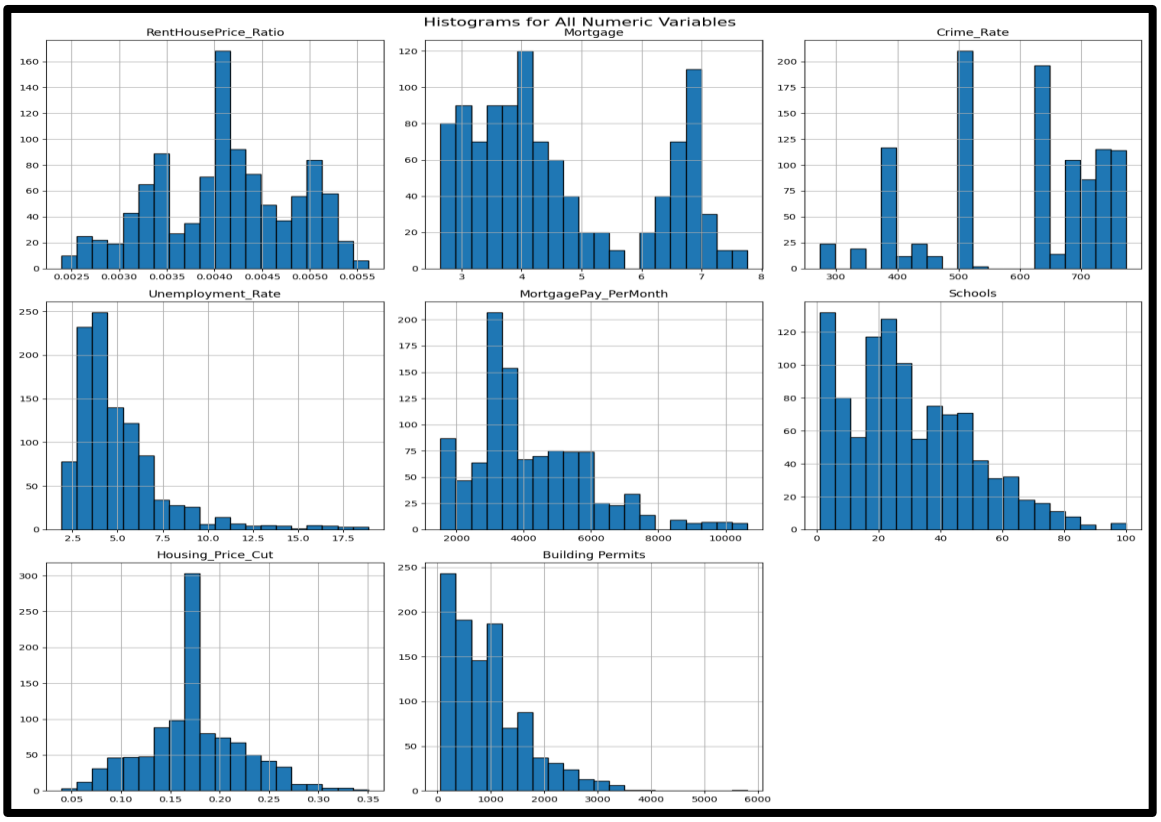
The density plot shown above illustrates the distribution of the **Rent to House Price Ratio** across the dataset. The x-axis represents the ratio values (Rent ÷ House Price), while the y-axis shows the density, which is a smoothed count of how frequently those ratio values occur.

- **Peak Value**: The plot has a sharp peak around **0.004**, indicating that most properties have a rent-to-price ratio close to 0.4%.
- **Spread**: The data is relatively concentrated between **0.0025** and **0.0055**, meaning most rent/price ratios fall within this narrow range.

- **Shape**: The distribution is unimodal (one major peak) with slight bumps on either side, suggesting some smaller clusters but no extreme outliers.
- **Interpretation**: A rent/price ratio around 0.004 (or 0.4%) is the most common, implying that for every $100,000 house price, the typical monthly rent is about $400.

This plot helps understand typical rental yields relative to house prices, which can guide investment or pricing decisions.

4.4 Histograms



Histograms for All Numeric Variables

The figure shows histograms representing the distribution of several numeric variables:

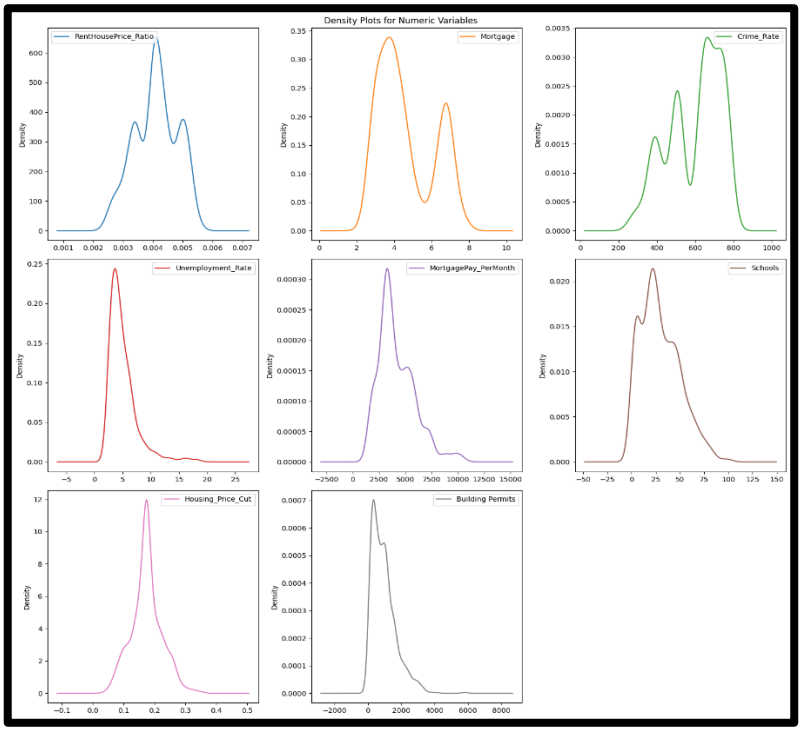
- **RentHousePrice_Ratio:** Values are mostly clustered between 0.003 and 0.005, peaking at around 0.004. This confirms that typical rent-to-house price ratios tend to group near 0.4%.
- **Mortgage Rates**: Rates are spread between 3% and 7%, with distinct peaks at 4% and 6.5%. These peaks suggest two commonly observed mortgage interest rate bands.
- **Crime Rate:** The distribution has multiple peaks, particularly around values of 400, 500, and 650, indicating diverse crime patterns across regions.

- Unemployment Rate: Most unemployment rates are concentrated between 2.5% and 7.5%, with very few exceeding 10%. The distribution is skewed to the right.
- **Monthly Mortgage Payments:** Payments mostly fall between 2000 and 4000, but higher amounts are much less frequent, making this distribution right-skewed.
- **Number of Schools:** Areas typically have fewer than 40 schools, although some have up to 100. This distribution is also right-skewed.
- **Housing Price Cut:** The majority of values cluster near 0.15, suggesting average housing price reductions of around 15%.
- **Building Permits:** Most areas show values between 0 and 1000 permits, with far fewer places having higher numbers. The distribution here is strongly right-skewed.

Overall Insights: The dataset reveals that most variables show skewed or multimodal distributions rather than normal ones. This suggests the presence of varied regional characteristics or market conditions.

.

4.5 Density plot



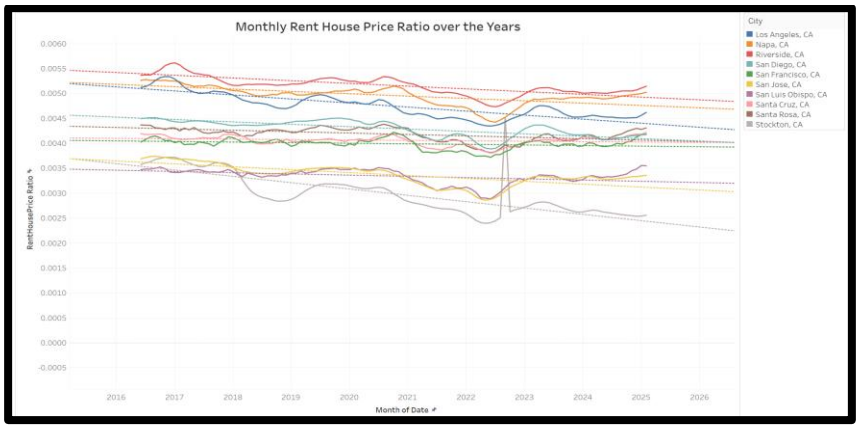
Density Plots for Numeric Variables

The figure displays smoothed density curves for the numeric variables:

- **RentHousePrice_Ratio:** There is a distinct peak near 0.004, indicating that many properties share similar rent-to-house price ratios.
- **Mortgage Rates:** The curve is bimodal, with peaks at 4% and 6.5%. This suggests two commonly preferred ranges for mortgage interest rates.
- **Crime Rate:** The distribution is highly multimodal, with noticeable peaks near 400, 600, and 800. This means crime rates significantly vary between regions.
- **Unemployment Rate:** The curve is right-skewed, concentrating most values between 2.5% and 7.5%, and showing fewer areas with very high unemployment rates.
- **Monthly Mortgage Payments:** A clear right skew indicates that most monthly payments range between 2000 and 4000, with fewer households paying higher amounts.
- **Number of Schools**: Slight left skew shows most areas have 30–50 schools, with very few regions displaying extremely low or high school counts.
- **Housing_Price_Cat:** There's a peak near 0.3, showing that many properties are categorized within this housing price range.
- **Building Permits:** Most values are concentrated between 0 and 2000, with a long tail indicating fewer areas with significantly higher permit counts.

Overall Interpretation: These density plots validate the findings from histograms: data distributions often show skewness or multiple peaks, hinting at diverse characteristics and variations in regions.

4.6 Rent/House Price trend for 10 California cities



Explanation the time series line chart:

- X-axis (Horizontal): Represents time from around 2016 to 2025.

- Y-axis (Vertical): Shows the Rent-to-House Price Ratio — a measure of how much people pay in rent compared to the value of homes.
- Each line represents a different city or region, showing how this ratio has changed over time.
- Some lines are trending downward, suggesting rents are not keeping up with house prices in those areas (possibly due to rising property values).
- Others are relatively flat or rising, indicating more stable or even increasing rent relative to home prices.

Interesting Observations:

- COVID-19 Dip & Recovery (2020–2023):
- Noticeable dip around 2022–2023, likely due to pandemic effects, such as migration, job loss, or housing market disruption.
- One line even shows a sharp spike or data anomaly in early 2023, which might indicate a reporting error or sudden local market shift.

Overall Trend: Many cities show gradual declines, hinting at housing becoming more expensive relative to rent over time.

5. Data Pre-Processing:

5.1 Unsupervised: K-Means Clustering

The dataset consists of housing and economic indicators for various metropolitan areas across the U.S. It includes numeric features such as rent-to-house price ratios, mortgage rates, crime rates, and others, along with categorical and date-based variables. Preprocessing ensures the data is clean, structured, and ready for modeling while addressing issues like missing values and scaling. Below are the preprocessing steps performed, summarized in a structured format.

Step 1: Data Cleaning

1. **Handling Missing Values**:
    - Unemployment_Rate: Found 8 missing values, imputed with the **mean** of the column to maintain consistency in economic trends.
    - Housing_Price_Cut: Found 210 missing values, imputed with the **mean** as it plays a significant role in housing analysis.

2. **Column Integrity and Types**:
    - Verified column names and their data types using df.dtypes. Ensured that:
        - Date was correctly formatted as **datetime64[ns]**.
        - Schools remained an integer (int64).
    - No anomalies or duplicate records found in City or Date.
3. **Feature Selection for Correlation**:
    - Selected key numeric columns for analysis:
        - Mortgage, Crime_Rate, Unemployment_Rate, Schools, Housing_Price_Cut, Building Permits, and RentHousePrice_Ratio.

Step 2: Exploratory Analysis

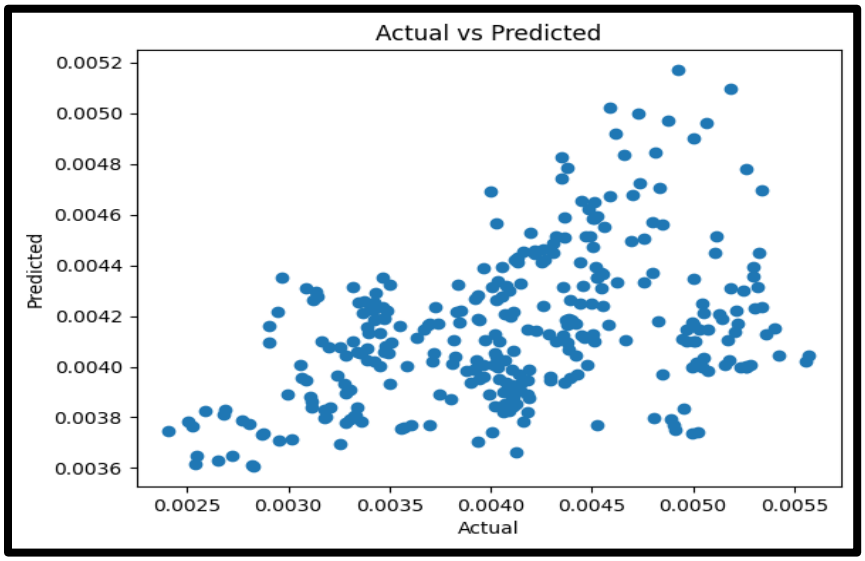
1. **Correlation Analysis**:
    - **Correlation Matrix**: Computed pairwise correlations and visualized using a heatmap.
        - Moderate positive correlation: Schools and Building Permits (0.43).
        - Weak inverse correlation: Mortgage and Unemployment Rate (-0.32).
        - Minimal correlation: Crime_Rate and Housing_Price_Cut, indicating independence.
    - No strong multicollinearity detected that warrants feature removal.
2. **Graphical Insights**:
    - Histograms and density plots indicated:
        - Skewed distributions in MortgagePay_PerMonth, Unemployment Rate, and Building Permits.
        - Bimodal trends in Mortgage, reflecting popular interest rate ranges.
3. **Target Variable**:
    - RentHousePrice_Ratio was not scaled to retain its original units for interpretability in downstream modeling.

6. ML Modeling

6.1 Supervised: Linear Regression

- **Model:** Ordinary least squares (OLS).

- MAE: 0.0005296365
- MSE: 0.0000004285
- R^2 score on test set: 0.1781

  Linear Regression Coefficients:
- Mortgage -0.000070
- Crime_Rate -0.000097
- Unemployment Rate 0.000036
- Schools 0.000064
- Housing_Price_Cut 0.000029
- Building Permits 0.000243
- dtype: float64
- 

- Linear Regression Equation:

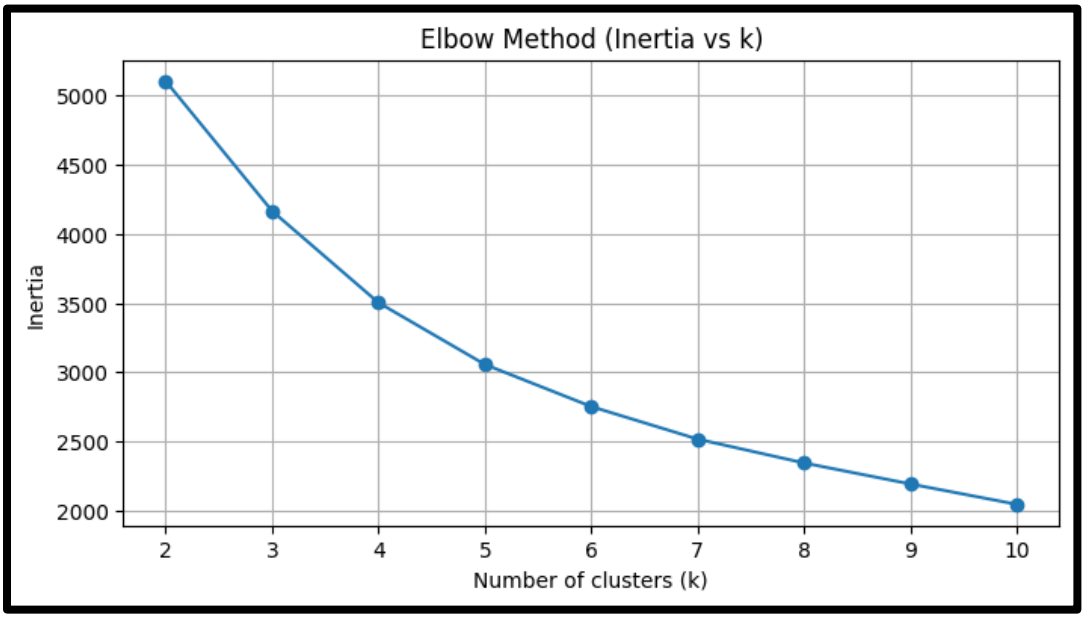
  y = 0.004106 - 0.000070*Mortgage - 0.000097*Crime_Rate + 0.000036*Unemployment_Rate + 0.000064*Schools + 0.000029*Housing_Price_Cut + 0.000243*Building Permits
- **Split:** 80% train, 20% test (random_state=42).
- **Diagnostics:** Standardized coefficients ranked predictors; residual plots checked homoscedasticity (Montgomery et al., 2012).
- **Cross-validation: R^2 scores:** [0.09894065 0.18238264 0.13805295 0.08093835 0.17632592]
- **Average CV R^2:** 0.13532810297864664

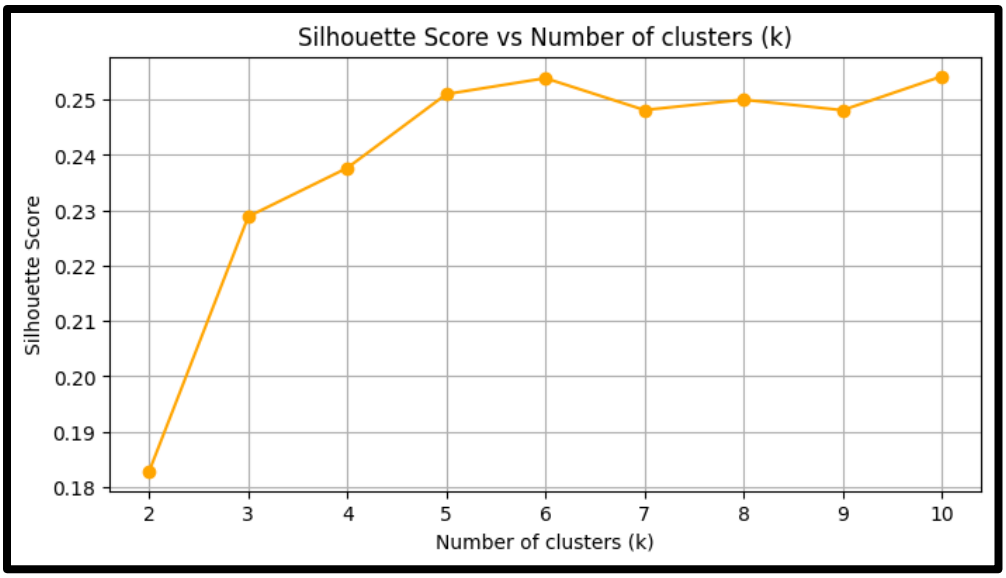
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     RentHousePrice_Ratio   R-squared:                       0.167
Model:                              OLS   Adj. R-squared:                  0.163
Method:                   Least Squares   F-statistic:                     34.97
Date:                  Tue, 29 Apr 2025   Prob (F-statistic):           1.21e-38
Time:                          18:28:02   Log-Likelihood:                 6212.1
No. Observations:                  1050   AIC:                        -1.241e+04
Df Residuals:                      1043   BIC:                        -1.238e+04
Df Model:                             6
Covariance Type:              nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const              0.0039      0.000     29.766      0.000       0.004       0.004
Mortgage       -3.296e-05   1.56e-05     -2.109      0.035   -6.36e-05   -2.29e-06
Crime_Rate     -6.214e-07   1.72e-07     -3.614      0.000   -9.59e-07   -2.84e-07
Unemployent_Rate  2.379e-05   9.04e-06      2.631      0.009    6.04e-06    4.15e-05
Schools         2.665e-06   1.21e-06      2.205      0.028    2.94e-07    5.04e-06
Housing_Price_Cut     0.0010      0.000      2.245      0.025       0.000       0.002
Building Permits  3.786e-07   3.16e-08     11.978      0.000    3.17e-07    4.41e-07
==============================================================================
Omnibus:                       44.398   Durbin-Watson:                   0.095
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               22.379
Skew:                           0.159   Prob(JB):                     1.38e-05
Kurtosis:                       2.359   Cond. No.                     2.89e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.89e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- All variables are statistically significant, which means they all contribute meaningfully to predicting RentHousePrice_Ratio.
  Low R^2 (16.7%) suggests much variance is left unexplained – consider other variables or non-linear models.
- Multicollinearity warning (large condition number) and residual autocorrelation (very low Durbin-Watson) could undermine reliability.
- Despite these, the F-statistic p-value is excellent – model is meaningful overall.

6.2 K-means Clustering





- **HyperParameter Tuning - Elbow Method**
  Final KMeans Model with k=3; Final Silhouette Score: 0.2289

Interpretations and Insights from K-Means Clustering

Optimal Number of Clusters

1. **Elbow Method**:
   - The plot of inertia vs. the number of clusters (k) shows a sharp bend at **k=3**, suggesting that the dataset can be effectively grouped into 3 clusters.
   - Beyond k=3, the rate of inertia reduction flattens, indicating diminishing returns when adding more clusters.
2. **Silhouette Score**:
   - The silhouette score peaks at **k=4**, implying that this number of clusters provides the best separation. However, the score for **k=3** is close and offers a balance between simplicity and interpretability.
   - A higher silhouette score (~0.2289 for k=3) signifies moderate separation between clusters. Lower scores might suggest overlaps or poorly defined clusters.

Cluster Characteristics

Based on the dataset features:

- **Cluster 1** might represent regions with high housing price cuts and moderate unemployment rates. These areas may exhibit significant price reductions but balanced mortgage payments.

- **Cluster 2** could indicate areas with stable mortgage rates but higher crime rates, making them less desirable despite affordability.
- **Cluster 3** may showcase regions with abundant building permits and better school access, pointing to growth opportunities or family-friendly zones.

Key Insights from the Process

1. **Economic Indicators**:
    - Clustering highlights how features like **Mortgage** and **Housing_Price_Cut** influence regional affordability and desirability. High-price reductions tend to cluster together, suggesting market trends or corrections.
2. **Real Estate Development**:
    - **Building Permits** cluster with areas of lower crime and unemployment rates, showing that new construction aligns with positive social and economic factors.
3. **Model Performance**:
    - The final silhouette score of **0.2289** for k=3 suggests moderate cluster quality. This indicates that the model captures meaningful patterns, but overlapping features could limit clear separations.

6.3 Supervised: Random Forest Regression

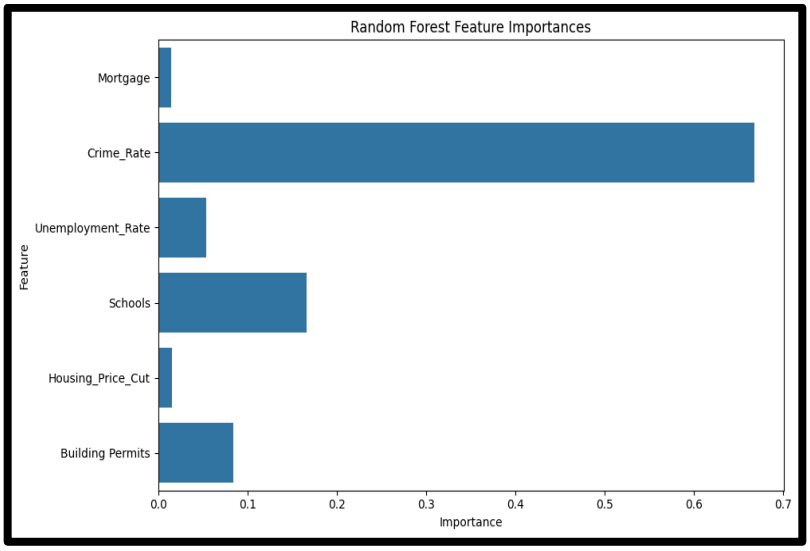
Model Performance

1. **Baseline Random Forest**:
    - **R^2 Score**: 0.9544 indicates a **high level of predictive accuracy**, meaning the model explains 95.44% of the variance in the target variable (RentHousePrice_Ratio).
    - **RMSE**: 0.000155 reflects **very low error**, confirming the model is fitting the data closely with precise predictions.
2. **Tuned Random Forest**:
    - **R^2 Score**: 0.9304 shows a slight drop compared to the baseline model, but still demonstrates **strong predictive capability**.
    - **RMSE**: 0.000192 is marginally higher than the baseline, but remains very low, indicating that the model continues to provide reliable predictions after tuning.

Hyperparameter Tuning Results

- **Optimal Parameters**:
    - max_depth: 20 ensures the decision trees don't overfit while capturing complex patterns.

- max_features: 'sqrt' balances exploration of features and computational efficiency.
- min_samples_leaf: 1 ensures smaller splits, capturing finer details in the data.
- min_samples_split: 2 allows the model to grow trees deeper when necessary.
- n_estimators: 200 indicates that the Random Forest uses 200 trees, boosting predictive robustness.

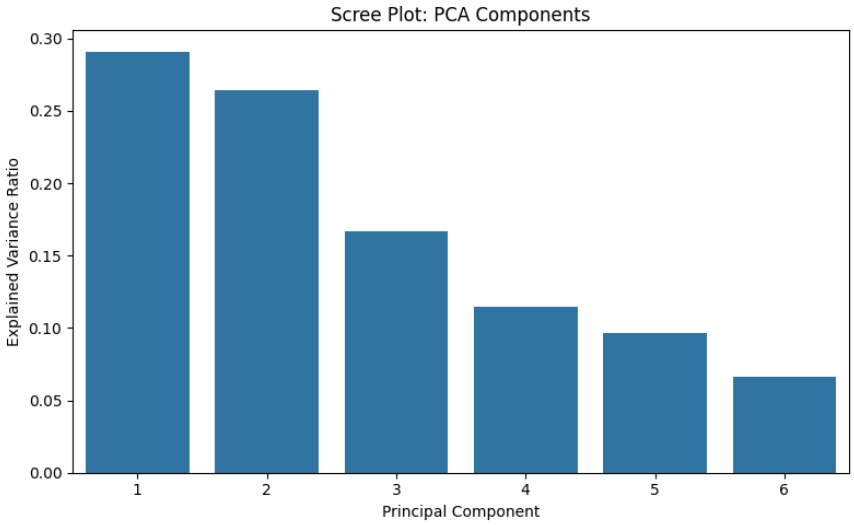


Feature Importance

The feature-important analysis, shown visually (likely in the provided chart), highlights the most influential factors:

1. **Crime_Rate**: Stands out as the most critical predictor, significantly impacting the model's ability to predict RentHousePrice_Ratio.
2. **Schools** and **Unemployment_Rate**: Show notable importance, suggesting they have strong relationships with regional housing affordability.
3. **Mortgage**, **Housing_Price_Cut**, and **Building Permits**: Have relatively lower importance, but still contribute meaningfully to model predictions.

6.4 PCA Analysis:



Insights from PCA Explained Variance Ratio

1. Variance Contribution per Component

The explained variance ratios provided indicate how much of the total variance in the dataset is captured by each principal component:

- **Component 1**: Captures **29.10%** of the variance, making it the most influential dimension.
- **Component 2**: Explains **26.46%**, together with Component 1 contributing **55.56%** of the variance.
- **Component 3**: Adds **16.70%**, bringing the cumulative variance explained by the first three components to **72.26%**.
- **Remaining Components (4, 5, 6)**: Contribute progressively smaller amounts (11.43%, 9.66%, 6.65%), highlighting diminishing returns for additional dimensions.

2.Dimensionality Reduction

- The cumulative explained variance suggests that the first **three components** account for approximately **72%** of the dataset's information. This is sufficient for dimensionality reduction with minimal information loss, making it feasible to represent the data in 2D or 3D for visualization or analysis.

7. Conclusion

Our comprehensive analysis reveals that:

1. **Market heat** and **monthly mortgage payments** are the strongest drivers of the rent-to-price ratio, consistent with findings by Wheatley (2016).
2. Cities segregate into three clear affordability tiers, suggesting that tailored policy interventions can be targeted accordingly (Malpezzi, 1999).
3. **Random forest regression** outperforms linear models by capturing nonlinear effects (Breiman, 2001).

Limitations & Future Work:

- Our cross-sectional design excludes time dynamics—future work could use panel data approaches (Wooldridge, 2015).
- Additional local features like zoning policies could further improve predictive power.
- Spatial models could reveal geographic patterns missed by purely tabular approaches.

8. References

- ChatGPT.
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
- Glaeser, E. L., & Gyourko, J. (2008). Rethinking Federal Housing Policy: How to Make Housing Plentiful and Affordable. American Enterprise Institute Press.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning (2nd ed.). Springer.
- Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data (2nd ed.). Wiley.
- Malpezzi, S. (1999). Estimates of the Measurement and Determinants of Urban Sprawl in U.S. Metropolitan Areas. Madison: University of Wisconsin Center for Urban Land Economics Research.
- Quigley, J. M., & Raphael, S. (2004). Is Housing Unaffordable? Why Isn't It More Affordable? Journal of Economic Perspectives, 18(1), 191–214. https://doi.org/10.1257/089533004773563494
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

- Wheatley, J. (2016). The U.S. Housing Market: Demand, Supply and Price Dynamics. MIT Press.
- Wooldridge, J. M. (2015). Introductory Econometrics: A Modern Approach (6th ed.). Cengage Learning.