Real Estate Price Analysis Report for Palm Beach County

(Prepared by: Sara Martin and Cinoo Bosco Thomas)

Abstract

This project investigates factors influencing the difference between list and sold prices in Palm Beach real estate. By analyzing property attributes, market conditions, and regional trends, the study uncovers actionable insights to refine pricing strategies, optimize sales, and improve transaction outcomes. The findings empower realtors and investors to make data-driven decisions. Future enhancements, such as advanced predictive models and additional features, can deepen insights, providing greater value to stakeholders.

Introduction

This analysis examines sold and rented properties in Palm Beach County during 2023 to identify factors driving disparities between listing and sale prices, as well as rental price trends. The study utilizes a dataset covering property characteristics such as bedrooms, bathrooms, square footage, lot size, year built, pool availability, and location.

Focusing on the question, **"Which factors most impact the difference between list and sold prices?"**, the research leverages data analytics to uncover patterns and correlations that explain these price variations. Key attributes analyzed include:

- Number of bedrooms (BEDS)
- Number of bathrooms (FBATHS)
- Square footage (TOT SQFT)
- Lot size (LOT SIZE)
- Year built (YEAR BUILT)

- Pool availability (POOL)
- Location (ZIP CODE)

Understanding these factors is crucial for sellers, buyers, and real estate agents to make informed decisions, optimize pricing strategies, and adapt to changing market dynamics.

Data Collection

The dataset used on this project was provided by the class instructor, and it contains real estate information for Palm Beach County mainly for the year 2023. The dataset includes 81 columns and 6,600 rows of which 3,400 were closed sales and 3,200 rented.

Data Collection and Preparation

Dataset Overview:

- Source: Provided by the class instructor.
- Size: 6,600 rows and 81 columns (3,400 closed sales, 3,200 rentals).

Key Variables:

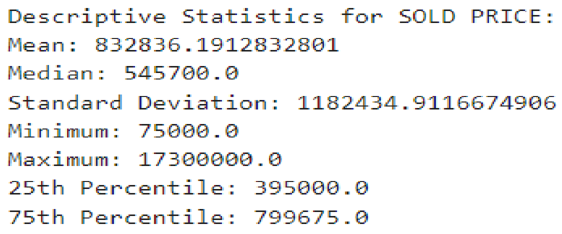
- Target Variable: Price_Difference = LIST PRICE - SOLD PRICE.
- Independent Variables: Bedrooms (BEDS), Bathrooms (FBATHS), Square Footage (TOT SQFT), Lot Size (LOT SIZE), Year Built (YEAR BUILT), Pool Availability (POOL), and Location (ZIP CODE).

Data Cleaning Steps:

1. Excluded rented properties to focus on sold properties.
2. Removed missing values in critical columns.
3. Encoded pool availability as a binary feature.
4. Calculated price difference (LIST PRICE - SOLD PRICE).
5. Standardized ZIP codes to analyze location-based variations.

Exploratory Data Analysis (EDA)

Descriptive Statistics: Summary statistics (mean, median, standard deviation) were calculated to understand the data distribution.



```
Descriptive Statistics for SOLD PRICE:
Mean: 832836.1912832801
Median: 545700.0
Standard Deviation: 1182434.9116674906
Minimum: 75000.0
Maximum: 17300000.0
25th Percentile: 395000.0
75th Percentile: 799675.0
```

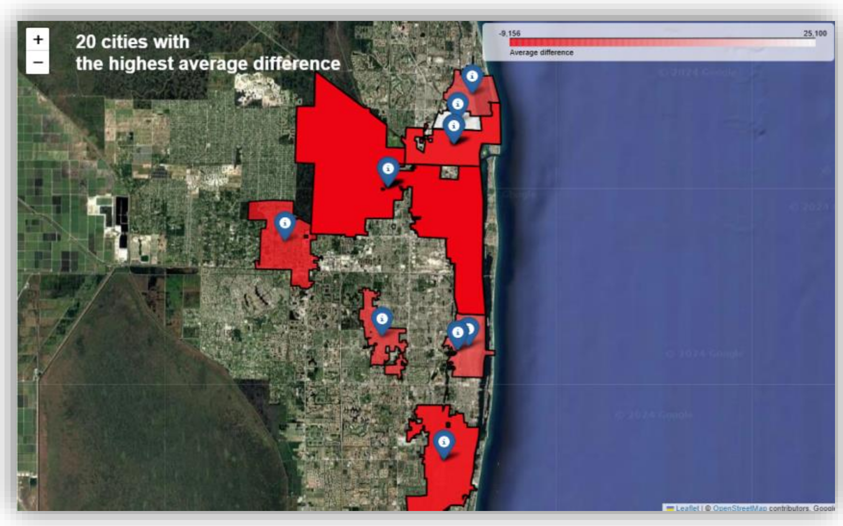
The **SOLD PRICE** data shows significant variability:

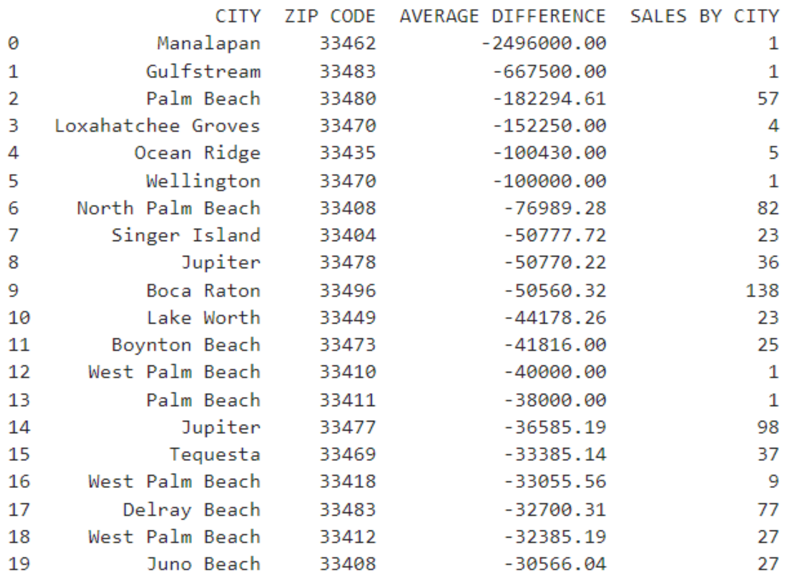
- **Mean**: 832,836, suggesting an average price, but it’s skewed by a few high-end properties.
- **Median**: 545,700, indicating that half of the properties sold for less, reflecting a right-skewed distribution.
- **Standard Deviation**: 1,182,434, indicating a wide variation in prices.
- **Minimum**: 75,000 and **Maximum**: 17,300,000, showing a large price range from low to luxury properties.
- **25th Percentile (395,000)** and **75th Percentile (799,675)** indicate that most properties fall within this price range.

Overall, the data shows a wide range of property prices with a skew towards higher-end properties.

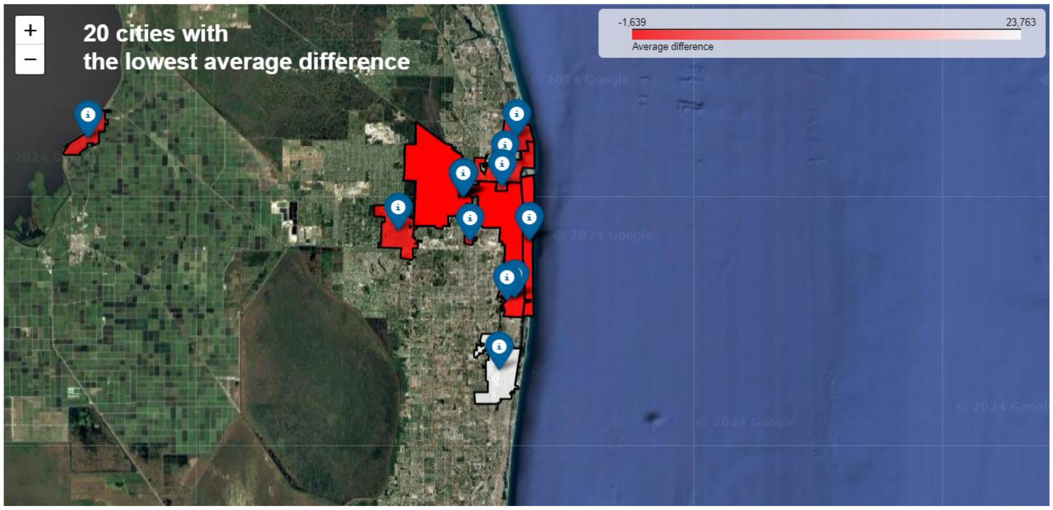
Market Trend Analysis

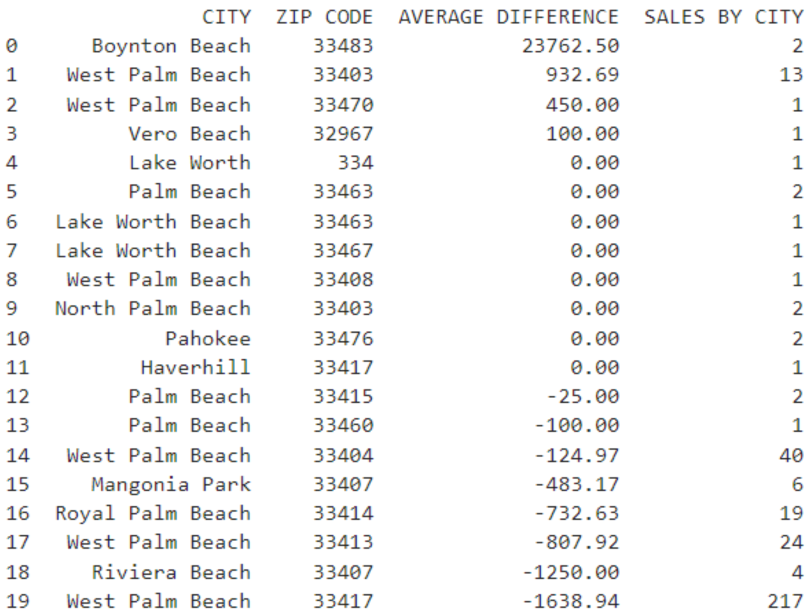
Cities with the highest and lowest average differences :


+
−
20 cities with
the highest average difference
-9,156
25,100
Average difference
Leaflet | © OpenStreetMap contributors, Googl


```
                  CITY  ZIP CODE  AVERAGE DIFFERENCE  SALES BY CITY
0            Manalapan     33462         -2496000.00              1
1           Gulfstream     33483          -667500.00              1
2           Palm Beach     33480          -182294.61             57
3    Loxahatchee Groves     33470          -152250.00              4
4          Ocean Ridge     33435          -100430.00              5
5           Wellington     33470          -100000.00              1
6      North Palm Beach     33408           -76989.28             82
7        Singer Island     33404           -50777.72             23
8              Jupiter     33478           -50770.22             36
9           Boca Raton     33496           -50560.32            138
10          Lake Worth     33449           -44178.26             23
11       Boynton Beach     33473           -41816.00             25
12     West Palm Beach     33410           -40000.00              1
13          Palm Beach     33411           -38000.00              1
14             Jupiter     33477           -36585.19             98
15            Tequesta     33469           -33385.14             37
16     West Palm Beach     33418           -33055.56              9
17        Delray Beach     33483           -32700.31             77
18     West Palm Beach     33412           -32385.19             27
19          Juno Beach     33408           -30566.04             27
```

The above data shows top 20 Cities with the highest average differences**.** These cities indicate markets where properties are often sold below the list price, potentially suggesting overpricing or a buyer's market. **Manalapan** leads with the largest loss (-$2.5M, 1 sale), followed by **Gulfstream** (-$667.5K, 1 sale), and **Palm Beach**, which had frequent sales (57) with an average loss of -$182K.



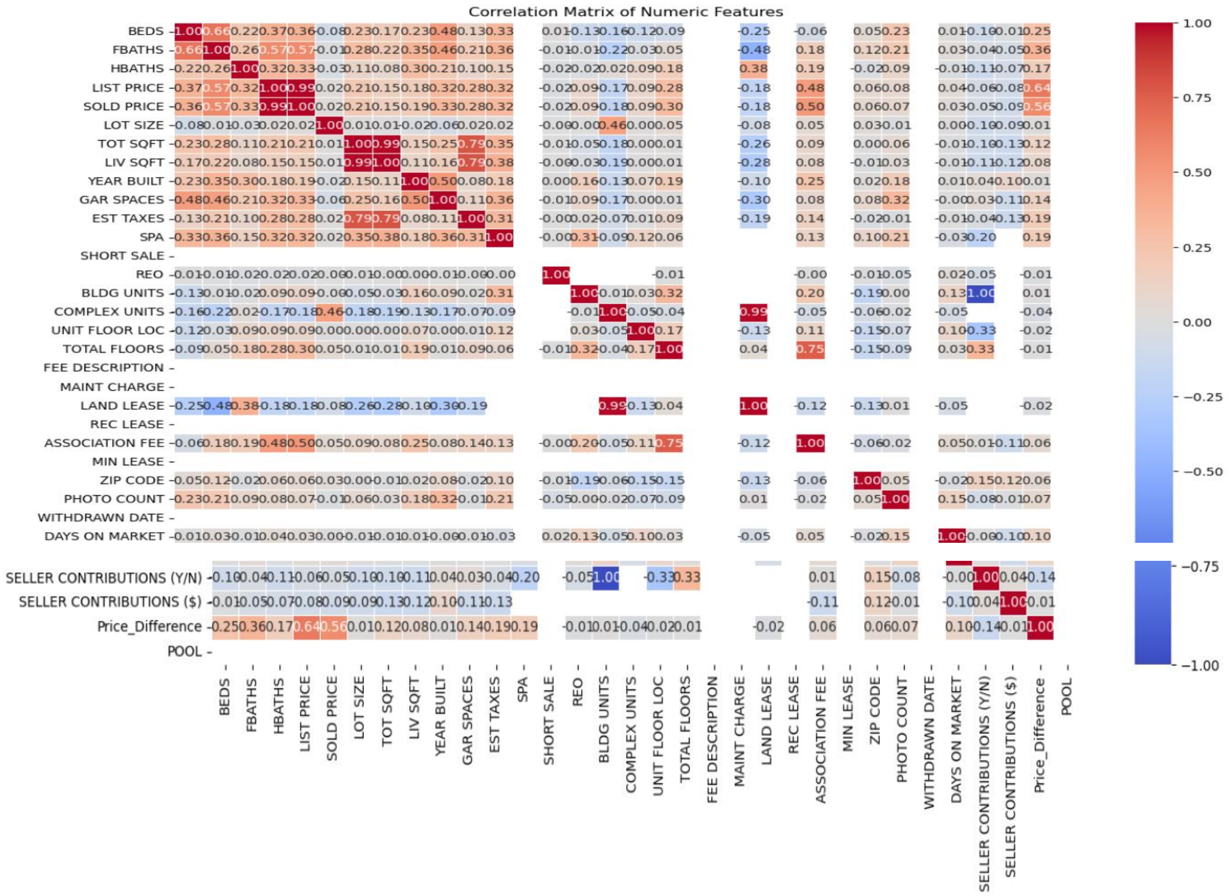


```
               CITY  ZIP CODE  AVERAGE DIFFERENCE  SALES BY CITY
0     Boynton Beach     33483            23762.50              2
1   West Palm Beach     33403              932.69             13
2   West Palm Beach     33470              450.00              1
3        Vero Beach     32967              100.00              1
4        Lake Worth       334                0.00              1
5        Palm Beach     33463                0.00              2
6  Lake Worth Beach     33463                0.00              1
7  Lake Worth Beach     33467                0.00              1
8   West Palm Beach     33408                0.00              1
9  North Palm Beach     33403                0.00              2
10          Pahokee     33476                0.00              2
11        Haverhill     33417                0.00              1
12       Palm Beach     33415              -25.00              2
13       Palm Beach     33460             -100.00              1
14  West Palm Beach     33404             -124.97             40
15    Mangonia Park     33407             -483.17              6
16 Royal Palm Beach     33414             -732.63             19
17  West Palm Beach     33413             -807.92             24
18    Riviera Beach     33407            -1250.00              4
19  West Palm Beach     33417            -1638.94            217
```

The above data shows top 20 cities with the lowest average differences. These markets reflect pricing stability or slight profit on sales, indicating efficient pricing or demand balance. **Boynton Beach** (+$23,762.50, 2 sales) and **West Palm Beach** (+$932.69, 13 sales) lead with positive differences. Many cities show minimal differences (0 to -100 USD), highlighting pricing stability near list values.

Correlation Analysis:

To better understand the relationships between numeric variables in the dataset, a correlation matrix was generated and visualized using a heatmap. This analysis highlights how different features interact and helps identify key drivers of price discrepancies



Key Observations

1. **Diagonal Values**: The diagonal values of the correlation matrix always equal 1.0, as these represent the correlation of each feature with itself. For example, the correlation of "BED" with "BED", "LIST PRICE" with "LIST PRICE", etc., will always be 1.0. These values serve as a baseline for comparison to see how other variables relate to each other.

2. **Strong Positive Correlations**: Positive correlations suggest that as property features (like size or amenities) increase, the price tends to rise accordingly.

- TOT SQFT and SOLD PRICE: A strong positive correlation (e.g., 0.79) indicates that larger properties tend to sell at higher prices.
- FBATHS and BEDS: A positive correlation (e.g., 0.66) suggests that homes with more bedrooms typically have more bathrooms, reflecting larger property sizes.
- LIST PRICE and SOLD PRICE: A very strong correlation (e.g., 0.99) shows that the listed price is a reliable predictor of the final selling price.

3. **Strong Negative Correlations**: Negative correlations show that as features like property age increase, selling prices tend to decrease.

- YEAR BUILT and SOLD PRICE: A negative correlation (e.g., -0.30) suggests that older homes tend to sell for less, likely due to maintenance needs or outdated features.
- TOTAL FLOORS and SOLD PRICE: A weak negative correlation (e.g., -0.19) indicates that multi-floor homes don't always sell for higher prices, as other factors like location or design may play a more significant role.

4. **Weak or No Correlations**: Features with weak correlations (close to **0**) contribute minimally to explaining price differences and are less useful for prediction compared to strongly correlated variables.

- POOL and SOLD PRICE: A low correlation (e.g., 0.05) indicates that having a pool has little influence on property selling prices, possibly due to maintenance concerns or regional preferences.
- SHORT SALE and SOLD PRICE: A near-zero correlation suggests that short sales don't consistently affect the final selling price

Models Used:

- **Linear Regression**: A simple model to understand linear relationships between features and the target variable.
- **Random Forest Regressor**: A non-linear model to capture complex patterns in the data.

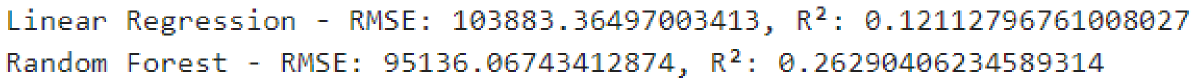
Model Training and Evaluation:

- The data was split into training (80%) and testing (20%) sets.
- Both models were evaluated using:

Root Mean Squared Error (RMSE): Measures the average difference between actual and predicted values.

R^2 Score: Indicates how well the model explains the variance in the target variable.

Results



```
Linear Regression - RMSE: 103883.36497003413, R²: 0.12112796761008027
Random Forest - RMSE: 95136.06743412874, R²: 0.26290406234589314
```

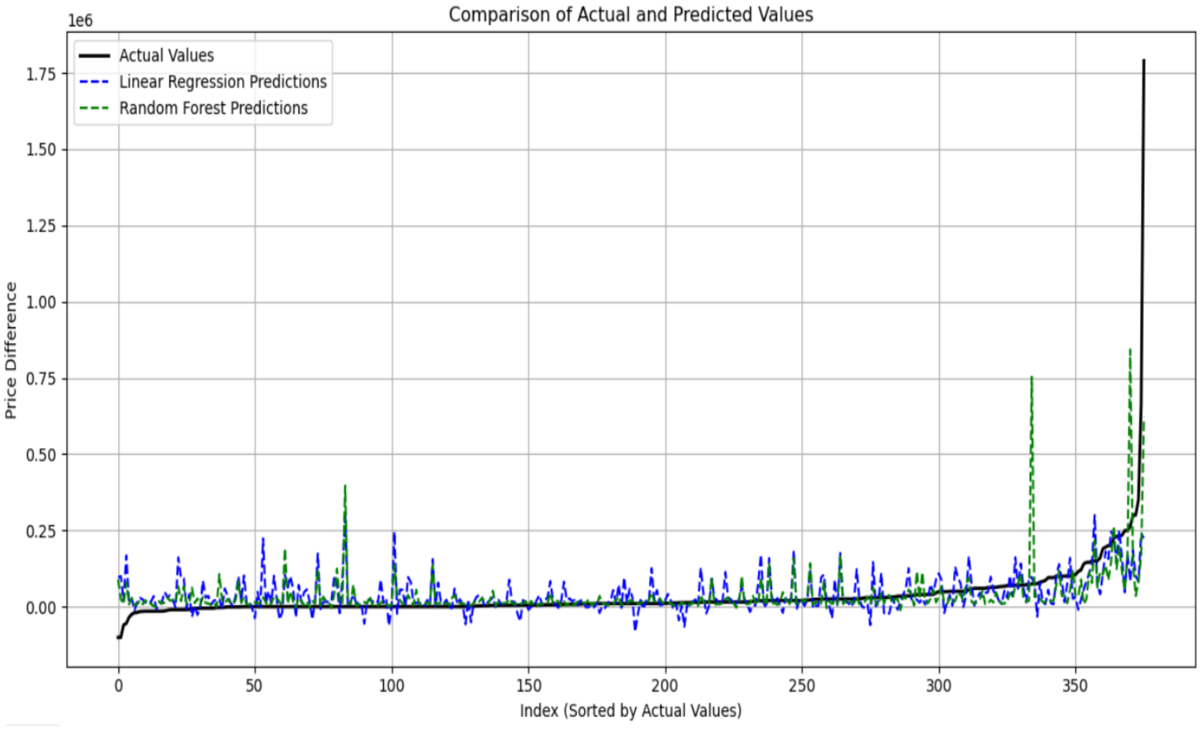
- **Linear Regression** has high RMSE (103,883) and low R^2 (0.121), indicating poor model fit and limited ability to explain the target variable.

- **Random Forest** performs better with lower RMSE (95,136) and higher R^2 (0.263), showing improved prediction accuracy and capturing more complexity in the data.

Overall, Random Forest outperforms Linear Regression, but both models still leave room for improvement.

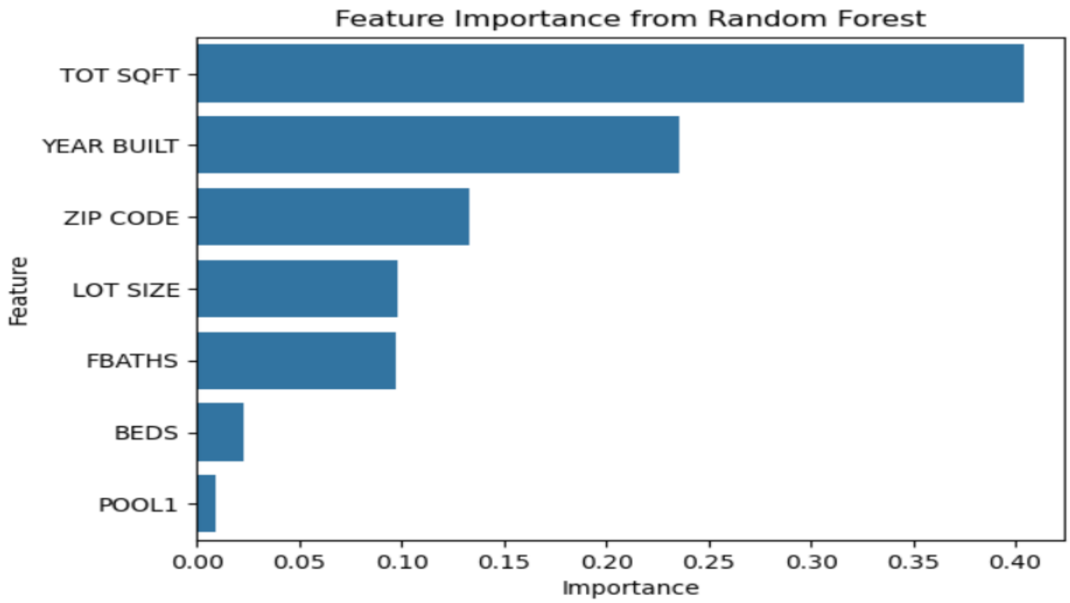
Model Comparison:

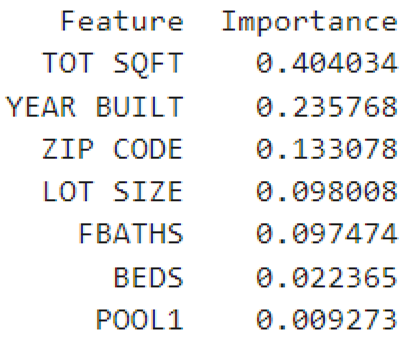
- A line plot compared predictions from Linear Regression and Random Forest against actual values.
- Random Forest predictions were closer to actual values.



Feature Importance:

The feature importances were calculated using the Random Forest model to determine the most influential factors. A bar plot was used to visualize these feature importances.





```
   Feature  Importance
  TOT SQFT    0.404034
YEAR BUILT    0.235768
  ZIP CODE    0.133078
  LOT SIZE    0.098008
    FBATHS    0.097474
      BEDS    0.022365
     POOL1    0.009273
```

The following features were identified as most impactful:

- **TOT SQFT:** Total square footage is the most important feature, contributing 40.4% to the model's prediction. This suggests that a larger square footage generally leads to a higher predicted price.
- **YEAR BUILT:** The year the house was built is the second most important feature, contributing 23.6% to the prediction. Newer houses are generally more valuable, so this feature likely has a positive impact on the price.
- **ZIP CODE:** T he zip code contributes 13.3% to the prediction. This indicates that location plays a significant role in determining home prices, likely due to factors like school districts, proximity to amenities, and neighborhood characteristics.
- **LOT SIZE:** Lot size contributes 9.8%, suggesting that larger lots are generally more valuable.

- **FBATHS:** The number of full bathrooms contributes 9.7%, indicating that having more bathrooms can positively impact the price.
- **BEDS:** The number of bedrooms contributes 2.2%, suggesting that while bedrooms are important, their impact is less significant compared to other features.
- **POOL1:** Whether the property has a pool contributes the least, at 0.9%. This suggests that having a pool has a relatively minor impact on the predicted price.

Overall, the model suggests that the total square footage, year built, zip code, and lot size are the most critical factors influencing home prices in this dataset.

Conclusions

Key Findings:

The dataset reveals that total square footage (TOT_SQFT), ZIP code (ZIP_CODE), and year built (YEAR_BUILT) were the primary factors influencing property sales in Palm Beach County in 2023. Despite most properties selling below their listed prices, buyers prioritize:

- **Larger Square Footage**: Offers more space and functionality for families and home offices, resulting in smaller differences between list and sold prices.
- **Year Built**: Homes constructed after 2005 meet higher construction standards, including better hurricane resistance, and are associated with smaller price discrepancies.
- **Location and property age**, especially ZIP codes in exclusive, safe neighborhoods, play a secondary but significant role in influencing property prices by driving higher appreciation over time.
- **The presence of a pool adds variability** to price differences, depending on buyer preferences and regional trends.

Model Recommendation:

The Random Forest Regressor is recommended for this task due to its better performance in capturing the complex relationships between features and the price difference.

Recommendations

- **For Sellers:** Accurately price properties with a market analysis, highlight square footage and bedrooms, and invest in modern upgrades and smart features.
- **For Buyers:** Negotiate aggressively in luxury markets, consider older properties for better deals, and focus on high-inventory areas with strong appreciation potential.

- **Future Research:** Include external factors (amenities, location scores) and use machine learning models to better predict price differences.

Contribution Statement

We, the project members, acknowledge that each member contributed equally to all deliverables, collaborating actively to ensure their success and coherence. Here is a breakdown of team members' contributions:

Sara Martin:

- Deliverable 2: Outlined Our Plan for Answering the Question (data collection process, variable types, definitions, and analysis plan)
- Deliverable 3: Created a detailed PDF describing each variable's purpose, data type, and context to support data interpretation and analysis.
- Deliverable 4: Handled data preprocessing, cleaning, and initial analysis of property differences. Created visualizations, optimized subplot layouts, interpreted chart trends, and analyzed geographic and temporal patterns to reveal meaningful insights.

Cinoo Bosco Thomas:

- Deliverable 2: Outlined Our Plan for Answering the Question (data collection process, variable types, definitions, and analysis plan)
- Deliverable 3: Provided a well-documented Python code snippet showcasing data manipulation, analysis, and visualization techniques to support project objectives.
- Deliverable 4: Trained the Random Forest model, tuned hyperparameters, and analyzed feature importances. Contributed to report writing, summarizing findings, and integrating visualizations and insights across all deliverables.