House Price Prediction Using PySpark

Abdulla Razick, Steven Sullivan, Uditi Shah
College of Computing & Informatics, Drexel University, Philadelphia, PA 19104, USA

1. Introduction

1.1 Project Scope and Objective

The House Price Prediction Project aims to develop a predictive model that estimates house prices based on historical data from various regions in the U.S. The primary dataset used for this project is the Zillow Home Value Index (ZHVI), which provides insights into housing price trends over time. This dataset spans multiple years and contains home values for different geographical regions, allowing us to analyze market fluctuations and trends.

The objective of this project is to build a machine-learning model that can accurately predict house prices using the ZHVI dataset. The model will leverage PySpark for data processing, exploratory analysis, and model development. By utilizing big data processing techniques, we aim to efficiently handle large-scale datasets and generate actionable insights for real estate market analysis.

1.2 Data Source

The data for this project is sourced from Zillow Research (Zillow Home Value Index (ZHVI)). The dataset includes monthly home value data starting from January 31, 2000, to December 31, 2024, across various regions in the U.S. The key columns in the dataset include:

- **RegionID** – Unique identifier for each region
- **SizeRank** – Rank of the region based on size
- **RegionName** – Name of the region (e.g., city or metropolitan area)
- **RegionType** – Type of region (e.g., metro, county, state, country)
- **StateName** – The U.S. state in which the region is located
- **Date Columns** – Home values recorded monthly over the years

1.3 Potential Users

- Real Estate Analysts: To predict and analyze future home prices in various regions.
- Investors: For better decision-making regarding home purchases or investments.
- Policy Makers: To understand trends and issues affecting housing markets.
- Homebuyers/Sellers: For more accurate estimations of home values.

2. Data Preprocessing

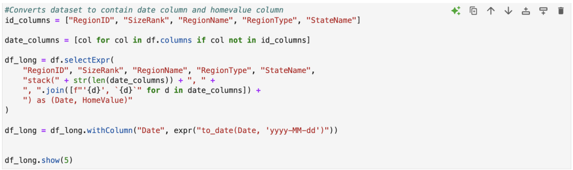
2.1 Overview

For this project, we utilized the Zillow Home Value Index (ZHVI) dataset and performed multiple data cleaning and transformation steps to ensure data quality. The key objectives of preprocessing were to handle missing values and engineer meaningful features.

2.2 Data Transformation

The dataset was transformed from wide format (multiple date columns) to a long format (single date column) to facilitate time-series analysis. This was achieved by melting the dataset, creating a structured table with:

- **RegionID, SizeRank, RegionName, RegionType, StateName**
- **Date** (transformed into a yyyy-MM-dd format)
- **HomeValue** (price data per region and date)

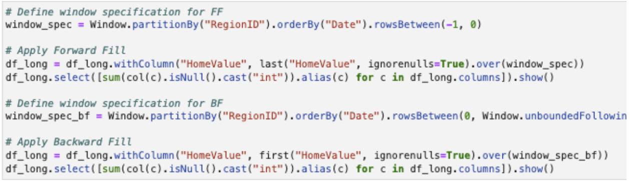


```
#Converts dataset to contain date column and homevalue column
id_columns = ["RegionID", "SizeRank", "RegionName", "RegionType", "StateName"]

date_columns = [col for col in df.columns if col not in id_columns]

df_long = df.selectExpr(
    "RegionID", "SizeRank", "RegionName", "RegionType", "StateName",
    "stack(" + str(len(date_columns)) + ", " +
    ", ".join([f"'{d}', `{d}`" for d in date_columns]) +
    ") as (Date, HomeValue)"
)

df_long = df_long.withColumn("Date", expr("to_date(Date, 'yyyy-MM-dd')"))


df_long.show(5)
```

2.3 Handling Missing Values

The dataset was initially examined for missing values across all columns. The primary columns with missing data were **HomeValue** and **StateName**.

- **HomeValue**: Forward Fill (FF) and Backward Fill (BF) were applied using PySpark Window functions to propagate the most recent available values
- **StateName**: Missing state names were filled by Assigning **"National"** to entries where **RegionName** was "United States".



```
# Define window specification for FF
window_spec = Window.partitionBy("RegionID").orderBy("Date").rowsBetween(-1, 0)

# Apply Forward Fill
df_long = df_long.withColumn("HomeValue", last("HomeValue", ignorenulls=True).over(window_spec))
df_long.select([sum(col(c).isNull().cast("int")).alias(c) for c in df_long.columns]).show()

# Define window specification for BF
window_spec_bf = Window.partitionBy("RegionID").orderBy("Date").rowsBetween(0, Window.unboundedFollowin

# Apply Backward Fill
df_long = df_long.withColumn("HomeValue", first("HomeValue", ignorenulls=True).over(window_spec_bf))
df_long.select([sum(col(c).isNull().cast("int")).alias(c) for c in df_long.columns]).show()
```

3. Exploratory Data Analysis (EDA)

3.1 Overview

The exploratory data analysis (EDA) phase focuses on understanding historical home price trends, regional variations, seasonal patterns, and key correlations within the dataset. This analysis provides foundational insights that guide the feature selection and modeling process.

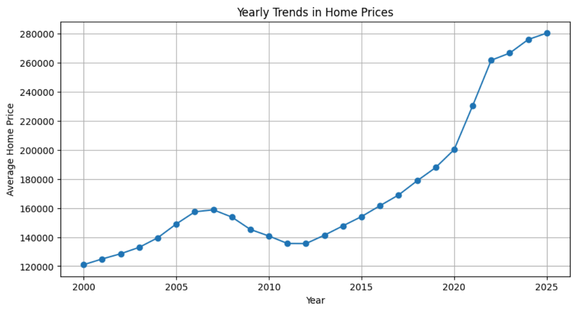
3.2 Missing Values Analysis

The dataset originally contained missing values in historical home price records. To address this, a two-step imputation strategy was employed:

1. Forward Fill (FF): Missing values were replaced with the most recent available home price for a given region.
2. Backward Fill (BF): Remaining missing values were filled using the next available price in the dataset.

Post-imputation verification confirmed the dataset is now complete and does not contain missing values, ensuring consistency in further analysis.

3.3 Yearly Home Price Trends

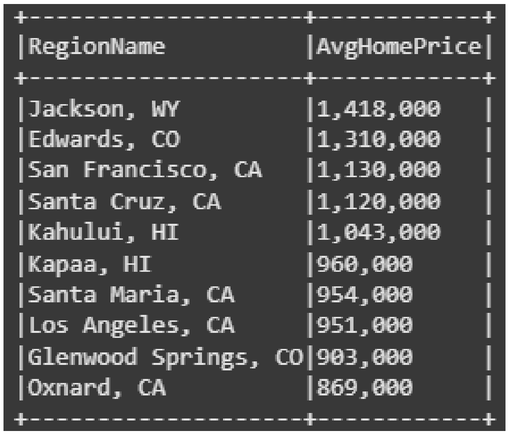


An analysis of historical home prices from 2000 to 2024 reveals a steady upward trend in property values. The data indicate that home prices remained relatively stable during the early 2000s but exhibited accelerated growth post-2015. This surge aligns with economic expansion, increased housing demand, and inflationary pressures. The observed trend suggests that real estate remains a strong long-term investment.

The observed increase in home values suggests a combination of rising demand, inflationary pressures, and supply constraints contributing to long-term price appreciation.

3.4 Most Expensive Regions

A regional analysis of the most expensive housing markets in the latest available year (2024) identifies Jackson, WY as the most expensive region, with an average home price of $1,418,000. Other high-value locations include Edwards, CO ($1,310,000), San Francisco, CA ($1,130,000), and Santa Cruz, CA ($1,120,000).



```
+--------------------+------------+
|RegionName          |AvgHomePrice|
+--------------------+------------+
|Jackson, WY         |1,418,000   |
|Edwards, CO         |1,310,000   |
|San Francisco, CA   |1,130,000   |
|Santa Cruz, CA      |1,120,000   |
|Kahului, HI         |1,043,000   |
|Kapaa, HI           |960,000     |
|Santa Maria, CA     |954,000     |
|Los Angeles, CA     |951,000     |
|Glenwood Springs, CO|903,000     |
|Oxnard, CA          |869,000     |
+--------------------+------------+
```

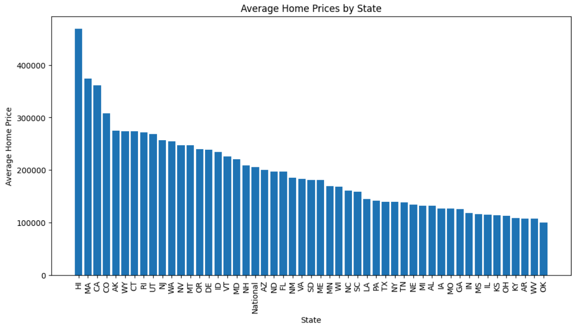
The table above presents the top 10 most expensive regions, showcasing the geographic distribution of high-value real estate markets. Notably, California dominates the list, with five cities appearing in the top 10. The presence of multiple Colorado and Hawaii cities suggests that luxury vacation destinations also contribute significantly to high housing prices.

Key Observations:

- California dominates the list, appearing five times in the top 10.
- Hawaii (Kahului and Kapaa) and Colorado (Edwards and Glenwood Springs) emerge as luxury real estate markets.
- The top cities share common factors, including strong job markets, tourism appeal, and limited housing supply.
- Jackson, WY tops the list, likely due to high demand for luxury properties and limited real estate availability.

These findings highlight the impact of location, economic activity, and desirability on home prices. The strong presence of coastal and tourist-driven markets indicates that demand in these areas remains resilient despite fluctuations in the broader real estate market.

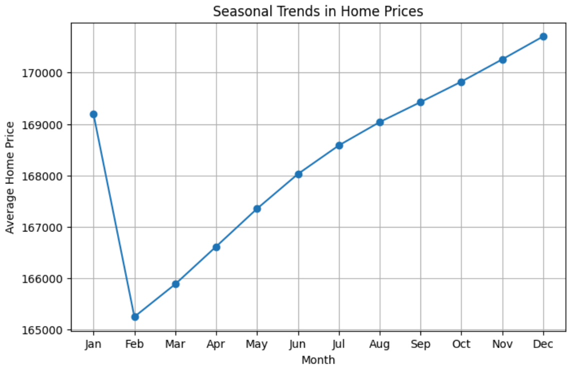
3.5 State-Level Home Prices



The state-level analysis further highlights substantial regional disparities in home values. California, Massachusetts, and Hawaii exhibit the highest average home prices, aligning with their strong economies, limited housing supply, and high desirability. Conversely, Midwestern and Southern states tend to have lower average home prices, likely due to lower population density and decreased housing demand.

The presence of high-value real estate markets in coastal and metropolitan areas aligns with expected trends, as these regions tend to attract higher economic activity and investment.

3.6 Seasonal Trends in Home Prices

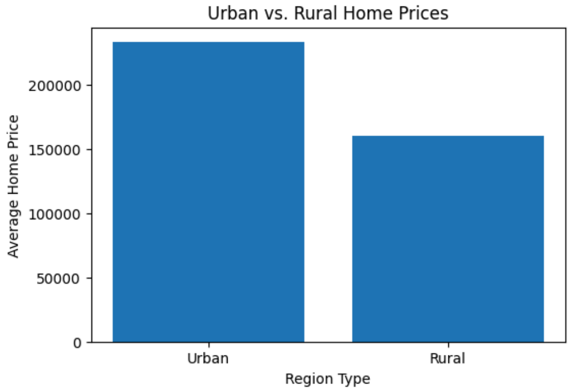


A seasonal analysis revealed that home values tend to peak during the summer months (June–August). This pattern can be attributed to:

- Increased buyer demand due to favorable weather conditions.
- Relocation timing based on school year transitions.
- Higher market activity, with more transactions occurring in summer.

This result is consistent with existing real estate market dynamics, where peak home-buying seasons drive up average prices.

3.7 Urban vs. Rural Price Comparison



A key aspect of the analysis was the comparison between urban and rural housing markets. Given the absence of explicit urban-rural classifications in the dataset, SizeRank was used as a proxy, with regions ranked 1-100 classified as urban and the rest as rural. The results indicate:

- Urban home prices are, on average, 45.36% higher than rural home prices.
- The price gap suggests that economic density, job opportunities, and housing demand drive urban home values higher.
- Rural markets remain more affordable, reinforcing the trend of lower housing demand in less densely populated areas.

These findings align with national housing trends, where high-demand metro areas drive up real estate prices, while rural regions remain cost-effective alternatives.

3.8 Correlation Analysis

A correlation analysis was performed to examine relationships between key features and home values. The results indicate:

- HomeValue & SizeRank (-0.25): A negative correlation suggests that larger metropolitan areas (lower SizeRank) tend to have higher home prices. This indicates that as a market size rank decreases, property values tend to increase, aligning with real estate market trends in densely populated cities.
- HomeValue & Year (+0.34): A positive correlation confirms the long-term appreciation of real estate, with prices increasing steadily over time. This result suggests that historical trends play a crucial role in predicting future home values.

These correlations confirm that urbanization and time-dependent factors are strong predictors of home values.

4. Machine Learning Models

Three machine learning models were implemented to predict home prices using the Zillow dataset:

1. **Linear Regression**:
   - A simple and interpretable model that assumes a linear relationship between features and the target variable (HomeValue).
2. **Random Forest Regression**:
   - An ensemble model that uses multiple decision trees to capture non-linear relationships and interactions in the data.
3. **Gradient Boosting Regression**:
   - A powerful ensemble model that builds trees sequentially, minimizing errors from previous trees.

Best Model

The **Gradient Boosting Regression** model is the **best-performing model** based on the following metrics:

- **Lowest RMSE (94468.23)**: Indicates the smallest prediction errors.
- **Lowest MAE (58686.07)**: Represents the smallest average absolute error.
- **Highest R² (0.2712)**: Explains the most variance in the target variable.
- **Best Visual Alignment**: Predicted vs. actual home values show the tightest clustering around the diagonal line.

Results

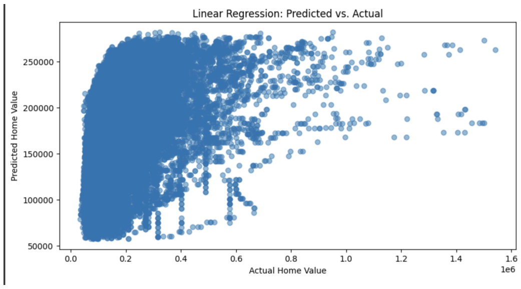
Model Performance Metrics

Model	**RMSE**	**MAE**	**R²**	**Training RMSE**	**Overfitting?**
Linear Regression	100285.47	63014.18	0.1787	100298.83	No
Random Forest Regression	94970.69	58910.90	0.2635	95194.22	No
Gradient Boosting Regression	94468.23	58686.07	0.2712	94624.46	No

Graph Explanations

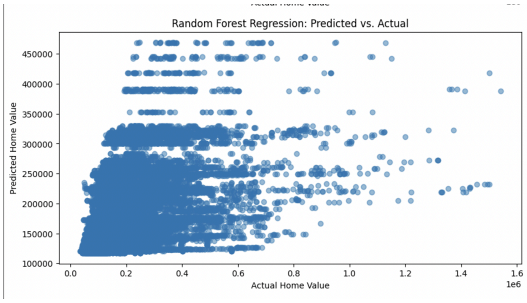
1. Linear Regression: Predicted vs. Actual

- The scatter plot shows **widely spread points**, indicating **poor prediction accuracy**.
- The points do not cluster tightly around the diagonal line (y = x), confirming that the model struggles to capture the underlying patterns in the data.
- This aligns with the **low R² value (0.1787)** and **high RMSE (100285.47)**.



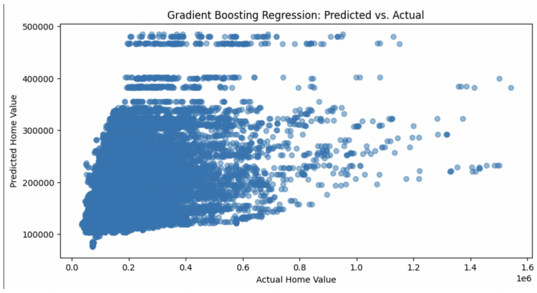
2. Random Forest Regression: Predicted vs. Actual

- The scatter plot shows **better clustering** of points around the diagonal line compared to Linear Regression.
- However, there is still some **spread**, especially for higher home values, indicating that the model struggles with extreme values.
- This aligns with the **moderate R^2 value (0.2635)** and **lower RMSE (94970.69)** compared to Linear Regression.



3. Gradient Boosting Regression: Predicted vs. Actual

- The scatter plot shows the **tightest clustering** of points around the diagonal line, indicating the **best prediction accuracy**.
- The points are more aligned with the diagonal, suggesting that the model captures the underlying patterns in the data well.
- This aligns with the **highest R^2 value (0.2712)** and **lowest RMSE (94468.23)**.



Conclusion

- **Gradient Boosting Regression** is the **recommended model** for predicting home prices due to its superior performance.
- **Random Forest Regression** is a good alternative if interpretability is important.
- **Linear Regression** is not suitable for this dataset due to its poor performance.

Expected Outcome

- The **Gradient Boosting Regression** model will provide the most accurate predictions of home prices, enabling users to:
  - Make informed buying/selling decisions.
  - Identify undervalued or overvalued properties.
  - Understand market trends and fluctuations.

Future Scope

1. **Feature Engineering**:
   - Add more relevant features (e.g., unemployment rate, interest rates, population growth) to improve model performance.
2. **Hyperparameter Tuning**:
   - Optimize hyperparameters for Gradient Boosting and Random Forest to further enhance accuracy.
3. **Geospatial Analysis**:
   - Incorporate geospatial data (e.g., proximity to schools, parks, transportation) to capture location-based trends.
4. **Advanced Models**:
   - Experiment with deep learning models (e.g., neural networks) for potentially better performance.