**DASC 5433 BIG DATA**

**URBAN DEVELOPMENT AND SUSTAINABILITY**

**Sustainable Cities and Communities: Big Data-Driven Analysis Using MapReduce**

**Instructor**

**Prof. Wu, Yalong**

**Team Members**

**Praneetha Bhattu,**

**Pratyusha Gadasanthala,**

**Naga Nikhitha Kondamuri,**

**Mohammed Zain Badal.**

**Table of Contents**

[Abstract 3](#_Toc195909619)

[1. Introduction 3](#_Toc195909620)

[1.1 Using this Project 4](#_Toc195909621)

[1.2 Prerequisite Knowledge for Beginners 4](#_Toc195909622)

[1.3 Problem Statement 4](#_Toc195909623)

[1.4 Objectives 4](#_Toc195909624)

[1.5 Relevance 5](#_Toc195909625)

[2. Literature Review 5](#_Toc195909626)

[2.1 Housing Prices and Sustainability 5](#_Toc195909627)

[2.2 Big Data Forecasting in Real Estate 5](#_Toc195909628)

[2.3 Why Further Research Is Needed 6](#_Toc195909629)

[3. Dataset Description 6](#_Toc195909630)

[3.1 Overview of the Dataset 6](#_Toc195909631)

[3.2 Data Preprocessing 7](#_Toc195909632)

[3.3 Exploratory Data Analysis (EDA) 9](#_Toc195909633)

[4. Methodology 10](#_Toc195909634)

[4.1 Techniques Used 10](#_Toc195909635)

[4.2 Design & Implementation 10](#_Toc195909636)

[5. Results 13](#_Toc195909637)

[5.1 Visualization 13](#_Toc195909638)

[5.2 Mapper & Reducer Code 17](#_Toc195909639)

[5.3 Parallel Processing 27](#_Toc195909640)

[6. Conclusion 32](#_Toc195909641)

[7. References 33](#_Toc195909642)

# **Abstract**

This report outlines the application of a big data analytics framework for analyzing urban housing and infrastructure trends using Hadoop and MapReduce. As global urbanization increases, cities face significant challenges including affordable housing shortages, slum proliferation, and infrastructure inadequacies. This project applies distributed computing to process large urban datasets, identify high-risk areas, and derive actionable insights for sustainable urban development. The solution employs MapReduce for data partitioning, cleaning, and trend analysis across key indicators like home type, year built, city demographics, and housing features.

**Keywords:** Big Data, Urbanization, MapReduce, Hadoop, Housing Analytics, Infrastructure Planning.

# **1. Introduction**

The global shift toward urban living has placed enormous pressure on cities to maintain livability, affordability, resilience, and inclusive infrastructure. As of the latest estimates by UN-Habitat, more than 55% of the world’s population lives in urban areas, a figure expected to rise to 68% by 2050. This accelerated urbanization has significant implications for social equity, environmental sustainability, and economic stability. Cities are now grappling with a multitude of issues, including inadequate and unaffordable housing, rising population density, unplanned slum growth, traffic congestion, insufficient access to public transportation, and overstretched public services. Traditional urban planning practices, often reliant on static surveys or local expertise, are increasingly inadequate to handle the volume, variety, and velocity of modern urban data.

Amid these challenges, big data technologies have emerged as powerful tools to support evidence-based urban governance. By leveraging vast amounts of structured and unstructured urban data, city administrators can gain insights into trends, identify problem areas, and deploy resources more strategically. However, the massive scale and fragmented nature of these datasets make conventional data analysis tools impractical. In this context, Hadoop and its core computational paradigm, MapReduce, offer a distributed and scalable solution to manage and analyze large-scale urban datasets efficiently.

This project centers on the design and implementation of a big data analytics framework aimed at promoting sustainable urban development through the use of MapReduce. Specifically, we analyze key urban indicators such as housing price trends, slum proliferation, infrastructure disparities, and housing configurations using real-world data. Our data sources include open government portals, UN-Habitat statistics, and public housing datasets. Through a systematic process involving data collection, cleaning, partitioning, and parallel analysis, we aim to uncover trends in affordability, infrastructure quality, and spatial inequality.

Our analysis includes four core modules: (1) comparing average house prices by city, (2) analyzing pricing trends by home type, (3) examining price variations based on house features such as number of bedrooms and stories, and (4) evaluating yearly housing price trends to understand market dynamics over time. The results of this analysis are intended to inform policymakers, urban planners, and community leaders by highlighting regional disparities and helping target investment and regulatory efforts.

Ultimately, the project showcases how Hadoop’s parallel computing capabilities can empower urban planning with real-time insights, foster data transparency, and contribute to the global Sustainable Development Goal 11: “Make cities and human settlements inclusive, safe, resilient and sustainable.”

## **1.1 Using this Project**

This project demonstrates how big data analytics can be effectively utilized to support sustainable urban development initiatives. By uncovering patterns in housing affordability, infrastructure disparities, and slum growth, the insights generated can be used by urban planners, government agencies, and NGOs to inform zoning regulations, infrastructure investments, and housing development policies. Furthermore, the modular approach used in the project enables scalability, allowing different cities to replicate the framework using their own datasets for tailored insights.

## **1.2 Prerequisite Knowledge for Beginners**

To fully understand this project and its applications, readers should have a basic understanding of:

* Urban development issues such as slums, housing affordability, and infrastructure planning
* Big data concepts including volume, velocity, and variety of data
* Core principles of Hadoop and the MapReduce programming model
* Fundamentals of data preprocessing, feature engineering, and categorical encoding Some familiarity with Python programming, CSV data manipulation, and distributed computing will help readers grasp the technical execution.

## **1.3 Problem Statement**

Urbanization has triggered numerous challenges for city planners, including inefficient housing allocation, rising inequality, and poor infrastructure development. Despite the increasing availability of urban data, most cities lack an integrated and scalable framework to harness this data for effective planning. Fragmented analysis of metrics such as housing cost, slum populations, and infrastructure coverage leads to inconsistent policy outcomes. Traditional analytical methods are ill-equipped to process such large and varied datasets, hindering timely and effective urban intervention.

## **1.4 Objectives**

The objectives of this project align with both sustainable urban development goals and the practical capabilities of big data processing using MapReduce. Our aim is to build a modular and scalable framework capable of analyzing housing and infrastructure patterns across large urban datasets. The following are the detailed objectives tied directly to the five mapper-reducer modules implemented:

1. **To assess housing affordability and pricing variations across cities** by processing large datasets of property sales. The city-wise mapper and reducer module computes the average house prices across multiple urban areas, enabling the identification of regions with severe affordability gaps.
2. **To analyze the impact of architectural design on pricing** using home type classification. By mapping different home types (e.g., townhomes, condos, single-family homes) against their sale prices, we provide insights into how design categories affect market value.
3. **To understand how interior features like the number of bathrooms, bedrooms, and stories influence pricing**, supporting developers and policy makers in optimizing home design for specific markets. The mapper-reducer for housing configurations helps quantify the pricing trends based on interior features.
4. **To track long-term real estate market trends over time** through year-wise aggregation of sale prices. This objective aids in understanding economic cycles, inflationary effects, and policy impacts on the housing sector.
5. **To explore how lot sizes vary with the number of bedrooms**, offering valuable insights into land-use planning and zoning efficiency. This analysis supports urban planners in identifying mismatches between house capacity and land allocation, which is key to managing urban sprawl.

Together, these objectives ensure a comprehensive analysis of urban housing data, addressing multiple dimensions of sustainability, affordability, infrastructure equity, and planning efficiency.

## **1.5 Relevance**

This project contributes directly to Sustainable Development Goal 11 by promoting data-driven approaches for building inclusive and sustainable urban environments. The methodology enables scalable insights across different dimensions of urbanization, from housing quality to infrastructure equity. As cities seek smarter ways to plan and grow, big data becomes a crucial asset in identifying challenges and designing actionable solutions. The framework presented in this project can help reduce slum populations, improve housing policy, and facilitate equitable infrastructure distribution—key aspects of long-term urban resilience and community well-being.

# **2. Literature Review**

Understanding the role of big data in urban development requires a careful review of existing housing market research and technology adoption literature. Two major strands of work emerge in the fields of housing affordability and data-driven real estate forecasting, each offering insights and identifying research gaps that this project addresses.

## **2.1 Housing Prices and Sustainability**

**Glaeser & Gyourko (2008)** explore housing affordability through a cost-based descriptive statistical approach. Their research highlights that housing is frequently overpriced in several urban areas, with high costs disproportionately affecting low- and middle-income families. However, their framework has limitations, particularly the lack of attention to long-term sustainability and the exclusion of non-economic dimensions such as environmental quality, equity, and community resilience. While their work laid the foundation for affordability research, it underscores the need for scalable, multidimensional, and future-facing housing analytics — a gap this project fills using big data.

## **2.2 Big Data Forecasting in Real Estate**

**YongLin Xiao (2002)** presents a different dimension of housing analysis through machine learning-driven big data modeling. The study applied artificial neural networks (ANN) and support vector machines (SVM) for predictive trend analysis in the real estate sector. Xiao emphasized the potential of these techniques in generating insights from large-scale property datasets, contributing to proactive market forecasting. However, the research also notes key barriers including high implementation costs, lack of standardized big data frameworks in real estate, and limited adoption outside major commercial hubs. These findings validate our choice of using Hadoop and MapReduce as cost-effective and scalable alternatives suitable for academic and public policy environments.

## **2.3 Why Further Research Is Needed**

Further literature suggests the urgency of new research due to the growing mismatch between housing demand and supply in rapidly urbanizing cities. Rising property prices, stagnating incomes, and insufficient planning contribute to reduced affordability and uneven access to urban opportunities. Thus, further research is needed to:

* Understand housing demand and supply trends through temporal and spatial data analysis.
* Develop better urban planning strategies by examining real-world housing configurations.
* Create affordability interventions through predictive and location-specific pricing patterns.

By integrating insights from these foundational studies and addressing their shortcomings through scalable and parallelized data analytics, this project contributes to a more nuanced, evidence-driven, and future-ready approach to sustainable housing policy.

# **3. Dataset Description**

## **3.1 Overview of the Dataset**

The cleaned dataset utilized in this study comprises **936,503 property records** and **23 structured columns**, offering a comprehensive snapshot of housing conditions, infrastructure availability, and transaction dynamics across urban areas. Each observation represents a unique housing unit, providing details about its physical characteristics, geographic location, sale history, and contextual surroundings such as local school ratings and property taxes.

The data was sourced from real estate aggregators and public housing datasets, capturing housing metrics across multiple U.S. cities. It reflects a diverse and representative cross-section of the housing market. With uniform formatting and low missing value frequency, this dataset is ideal for distributed computation using Hadoop's MapReduce model. Its size and variability allow for high-resolution analysis of housing inequality, price behavior, and urban sustainability trends.

A screenshot of a computer

AI-generated content may be incorrect.

**3.1.1 Key Features of the Dataset:**

* **Locational Attributes:** city, zipcode — These allow spatial grouping of housing trends and comparisons between regions.
* **Structural Features:** garageSpaces, parkingSpaces, numOfBathrooms, numOfBedrooms, numOfStories, livingAreaSqFt, lotSizeSqFt, homeType, yearBuilt — These represent the core characteristics of each property.
* **Utility/Condition Indicators:** hasCooling, hasGarage, hasHeating — Binary indicators showing whether a property includes key utility features.
* **Transaction and Pricing:** latestPrice, latestPriceSource, numPriceChanges, latest\_saledate, latest\_salemonth, latest\_saleyear — These variables capture market activity and valuation over time.
* **Educational Context:** numOfElementarySchools, avgSchoolRating — Help assess the impact of nearby school infrastructure on property valuation.
* **Tax-Related:** propertyTaxRate — Offers insights into local tax burdens and affordability barriers.

**3.1.2 Significance of the Dataset:** This dataset provides a multidimensional view of urban housing markets, ideal for examining disparities in affordability and infrastructure access. It supports spatial, temporal, and categorical analysis — enabling meaningful interpretation of patterns such as housing distribution by home type, property size, school ratings, and tax rates. When processed in a MapReduce environment, this dataset allows scalable computations to uncover city-level and feature-level trends, forming the analytical core for policy decisions and urban planning aligned with **Sustainable Development Goal 11**: “Make cities and human settlements inclusive, safe, resilient, and sustainable.”

## **3.2 Data Preprocessing**

**3.2.1 Data Standardization and Cleaning:**

* Removed non-contributory and redundant columns including zpid, streetAddress, latitude, longitude, multimedia descriptors, and other highly sparse fields.
* Harmonized categorical fields and standardized Boolean columns (e.g., hasCooling, hasGarage) for compatibility with MapReduce.
* Minimal null values (1 per column) were cleaned through direct row removal due to large sample size and minimal loss of data.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 1: Missing values per column before cleaning — Each column had only 1 missing value, supporting direct row-level removal.**

**3.2.2 Outlier Detection**

* Applied rule-based filters:
  + **Removed low-priced homes** (< $50,000), often representing distressed or misrecorded listings.
  + **Excluded ultra-luxury properties** (> $1.2 million), to reduce skewness in aggregate computations.
* This step ensured more representative analysis of mid-income and typical urban housing.

A comparison of a number of bars

AI-generated content may be incorrect.

**Figure 2: Effect of outlier removal on price distribution**

**3.2.3 Feature Engineering**

* age\_of\_house: Derived as 2025 - yearBuilt, useful for depreciation modeling and trend analysis.
* price\_per\_sqft: Standardized price by size to allow comparison across diverse property types.
* latestPrice\_log: Created using log1p() transformation to reduce skew and stabilize variance for model inputs.
* Applied **MinMaxScaler** on numerical columns like lotSizeSqFt, propertyTaxRate, and avgSchoolRating to prepare for uniform range-based processing and visualizations.

## **3.3 Exploratory Data Analysis (EDA)**

To understand the structure and distribution of the cleaned housing dataset, a comprehensive Exploratory Data Analysis (EDA) was performed. This phase aimed to identify relationships between key housing attributes—such as home type, bedroom count, sale year, and location—and their influence on price and demand. The results helped guide the logic of MapReduce design and supported the interpretation of high-dimensional urban housing trends.

**3.3.1 Visualization Techniques**

Multiple visual methods were used to explore patterns in the data:

* **Scatter Plot: Average Lot Size vs. Number of Bedrooms**: Visualized the relationship between bedroom count and land usage, indicating how property size scales with lot area.
* **Bar Chart: Average Home Price by Property Type**: Compared the mean property value across various home types (e.g., Condo, Single Family, Townhouse, Multifamily).
* **Line Graph: Average Sale Price by Year (2000–2024)**: Displayed historical pricing trends to identify long-term stability or economic disruptions in the housing market.
* **Horizontal Bar Chart: Top 10 Austin Neighborhoods by Sales Volume**: Ranked neighborhoods based on total transactions to highlight demand for hotspots within the city.
* (Pending) **Graph 5: Price Variation by Structural Layout (Bedrooms, Bathrooms, Stories)**: Will analyze how specific layout configurations impact pricing and frequency of sales.

**3.3.2 Key Insights**

* **Property Size Scales with Bedrooms**: The scatter plot showed a positive trend—properties with more bedrooms consistently occupied larger lot sizes. This confirms the expected zoning relationship between house size and land allocation.
* **Single Family Homes Are the Most Affordable**: Among all home types, Single Family homes had the lowest average price, while denser residential options (Condos, Townhouses, Multifamily units) held higher valuations, suggesting stronger urban demand.
* **Stable Long-Term Pricing with Temporary Dip**: Housing prices remained stable for nearly two decades, with a sharp decline in 2019–2020 likely due to external factors such as COVID-19 or reporting gaps. Prices recovered quickly, indicating market resilience.
* **High-Sales Neighborhoods Reveal Market Activity Clusters**: Neighborhoods like Port Michael and West Michael recorded the highest transaction volumes, marking them as high-turnover zones. These findings are valuable for real estate targeting and municipal development planning.
* **Foundation for Scalable Analysis**: These EDA results directly informed the design of the five MapReduce tasks, influencing how data was partitioned, grouped, and analyzed at scale.

# **4. Methodology**

## **4.1 Techniques Used**

**4.1.1 Big Data Tools**

* **MapReduce:** Employed to perform parallel processing on the cleaned housing dataset. The framework splits large tasks into smaller, distributed computations via custom **Mappers** and **Reducers**. Each mapper emits key-value pairs (e.g., city, price), while reducers perform aggregations (e.g., average, count) based on grouped keys.
* **Hadoop:** Served as the core framework for scalable and fault-tolerant processing. Hadoop’s architecture enables data locality, parallel execution, and seamless integration of custom Python scripts using **Hadoop Streaming**.

**4.1.2 Data Partitioning**

* Partitioning was **dynamically managed by each mapper**, rather than pre-splitting the dataset. Each MapReduce job focused on a specific analytical dimension:
  + city → for geographic pricing patterns
  + homeType → for home category trends
  + yearBuilt, bedrooms, lotSizeSqFt → for structural and temporal comparisons
* This flexible, attribute-driven partitioning allowed parallel execution of analytical tasks without hard-coded data slicing, enhancing scalability across housing attributes.

**4.1.3 Distributed Storage**

* **HDFS (Hadoop Distributed File System):** The cleaned dataset was stored in HDFS and split automatically into fixed-size blocks (default 128 MB). These blocks were distributed across the Hadoop cluster.
* **Data Replication:** HDFS maintained **multiple replicas** of each block (typically 3) to ensure data availability and fault tolerance even if one or more nodes failed.
* **Parallel Access:** Mappers were executed locally on the nodes holding the relevant HDFS blocks, reducing network transfer overhead and accelerating task performance.

## **4.2 Design & Implementation**

**4.2.1 Data Collection and Ingestion**

The dataset was compiled from public housing data portals and real estate aggregators. It covered:

* 936,503 housing entries across various U.S. cities
* 23 structured attributes ranging from structural details (bedrooms, livingAreaSqFt) to environmental factors (avgSchoolRating) and transactional metadata (latestPrice, saleYear).

Data was preprocessed in Python, saved as a CSV file, and ingested into HDFS using: hdfs dfs -put cleaned\_data.csv /input/

**4.2.2 Data Preprocessing and Cleaning**

Preprocessing was performed prior to Hadoop ingestion to reduce computational burden:

* **Dropped 20+ sparse or irrelevant columns** (e.g., zpid, streetAddress, numOfPhotos)
* **Handled Missing Values:** Each column had exactly 1 null entry due to a single corrupted row, which was removed.
* **Filtered Outliers:**
  + Lower threshold: Homes priced < $50,000
  + Upper threshold: Homes priced > $1.2 million
* **Feature Engineering:**
  + price\_per\_sqft = latestPrice / livingAreaSqFt
  + age\_of\_house = 2025 - yearBuilt
  + latestPrice\_log = log1p(latestPrice)

This ensured a clean, numerically stable, and analysis-ready dataset for distributed processing.

**4.2.3 Data Partitioning and Distribution**

In this project, data partitioning was aligned directly with the design of each custom **MapReduce task**, rather than being pre-defined based on static fields. The full cleaned dataset was passed as input to Hadoop, and partitioning was **functionally driven by mapper logic**, with each mapper responsible for extracting and grouping data by a specific attribute.

Each of the five mapper-reducer modules served a unique analytical goal and internally handled grouping based on one of the following keys:

* **City Mapper:** Partitioned the dataset by city to compare average housing prices across geographic regions.
* **Home Type Mapper:** Grouped data by homeType (e.g., Single Family, Condo) to identify pricing disparities across property types.
* **Features Mapper:** Segmented homes by structural features such as number of bedrooms, bathrooms, and stories to analyze valuation differences by layout.
* **Year Built Mapper:** Organized records by yearBuilt to explore temporal trends in property pricing over construction eras.
* **Bedrooms vs. Lot Size Mapper:** Grouped by numOfBedrooms to examine how land allocation varies by bedroom count.

This flexible, mapper-centric approach to partitioning enabled highly **targeted parallelism**, with each mapper emitting key-value pairs on-the-fly and Hadoop automatically routing similar keys to the same reducer.

Unlike traditional static partitioning schemes, this **dynamic partitioning model** leveraged MapReduce’s shuffling and sorting phase to achieve distributed grouping — optimizing processing efficiency while supporting modular analysis across multiple housing dimensions.

**Figure 3: Workflow**

The flowchart illustrates the end-to-end pipeline of this housing price prediction project, starting with the acquisition of over 936,000 property records from real estate portals and public housing datasets. These records include diverse housing attributes such as city, home type, square footage, tax rates, school ratings, and sale history.

The raw dataset undergoes extensive preprocessing using Python and Pandas. This includes the removal of sparse or irrelevant columns, elimination of a single incomplete row, and the transformation of key variables. New features such as price\_per\_sqft, age\_of\_house, and log-transformed price fields are engineered to improve data quality and analytical performance.

The cleaned dataset is then uploaded to the Hadoop Distributed File System (HDFS), where MapReduce tasks are executed in parallel across the cluster. Each mapper categorizes data by key attributes such as city, home type, year built, or number of bedrooms—emitting structured key-value pairs to downstream reducers. Hadoop's shuffle and sort mechanism groups identical keys, ensuring that reducers receive correctly partitioned data segments.

Each reducer computes aggregated statistics such as average prices, total sales, or feature-specific trends. For example, the project calculates the average price per home type, lot size distributions by bedroom count, and price changes across construction years.

Finally, the results are collected, visualized, and analyzed to uncover meaningful patterns in urban housing equity. These insights support sustainable city planning, affordability analysis, and targeted urban interventions, aligning with the goals of SDG 11: Sustainable Cities and Communities.

**4.2.4 Mapper Design**

Each Mapper reads through the dataset to extract and emit key-value pairs based on a specific housing attribute relevant to the analysis:

* **City Mapper:** Extracts city and latestPrice. Emits key-value pairs in the format: city \t latestPrice
* **Home Type Mapper:** Extracts homeType and latestPrice. Emits: homeType \t latestPrice
* **Structural Features Mapper:** Extracts numOfBedrooms, numOfBathrooms, and numOfStories. Emits: numOfBedrooms,numOfBathrooms,numOfStories \t latestPrice
* **Year Built Mapper:** Extracts yearBuilt and latestPrice. Emits: yearBuilt \t latestPrice
* **Lot Size Mapper:** Extracts numOfBedrooms and lotSizeSqFt to explore how lot size varies with bedroom count. Emits: numOfBedrooms \t lotSizeSqFt

Each mapper processes the full dataset and filters out missing or malformed records to ensure clean key-value output for reducers.

**4.2.5 Reducer Design**

Each Reducer aggregates the intermediate key-value pairs received from Mappers and computes summary statistics:

* **City Reducer:** Calculates average housing price per city, enabling geographic comparison of affordability.
* **Home Type Reducer:** Computes average price per homeType to reveal trends across property categories like condos, single-family, and townhomes.
* **Structural Features Reducer:** Aggregates prices by combinations of bedrooms, bathrooms, and stories to identify popular housing layouts and their valuation trends.
* **Year Built Reducer:** Calculates average price by yearBuilt, helping analyze how construction era influences property valuation.
* **Lot Size Reducer:** Computes average lot size for each bedroom count to identify land distribution trends in residential zones.

All reducers write results in tab-delimited text format, with keys as grouping variables and values as the aggregated metric (e.g., average, count).

**4.2.6 Execution in Hadoop Framework**

The project’s MapReduce scripts are executed within the Hadoop ecosystem as follows:

* **Mappers** run in parallel across distributed nodes, reading the HDFS-stored dataset block-by-block.
* The **Shuffle and Sort phase** organizes the emitted key-value pairs so that all values associated with a key are grouped together.
* **Reducers** then process these grouped values to compute statistical summaries, such as average home price per city or typical lot size for 3-bedroom houses.
* The final outputs are written to HDFS output directories and merged for post-processing.

# **5. Results**

## **5.1 Visualization**

This section showcases key visualizations generated from the results of the MapReduce tasks. Each figure offers a visual summary of patterns across pricing, structure, and geography, enabling quick interpretation of complex housing trends. These visualizations are instrumental in identifying market behaviors and guiding sustainable urban development initiatives.

**a. Average Lot Size vs. Number of Bedrooms:** This scatter plot illustrates how the average lot size (measured in square feet) varies with the number of bedrooms in a property. The x-axis represents the number of bedrooms, ranging from 1 to 20, while the y-axis reflects the corresponding average lot size for each bedroom category.

**A graph with blue dots

AI-generated content may be incorrect.**

**Figure 4: Average Lot Size vs. Number of Bedrooms**

**Interpretation & Insight:** The output displays a clearly **positive linear trend**—as the number of bedrooms increases, the average lot size also increases. This is expected in residential zones, where homes with more rooms typically require larger plots of land. Notably:

* A sharp upward curve is observed beyond 8 bedrooms, suggesting that very large homes (often luxury estates) occupy significantly more space.
* The outlier at 20 bedrooms reinforces this pattern, reaching the highest average lot size.

This visualization validates the correlation between property size and land usage, helping inform zoning guidelines and urban density strategies.

**b. Average Home Price by Property Type:** This bar chart compares the average housing price across different home types such as Single Family, Condo, Townhouse, and Multifamily. Prices are displayed in millions of dollars.

**A graph showing a number of houses

AI-generated content may be incorrect.**

**Figure 5: Average Home Price by Property Type**

**Interpretation & Insight:** The output displays significant **variation in pricing** across home types:

* **Single Family homes** are the most affordable option, with an average price of $4.46M.
* Other property types—such as **Condos, Townhouses, Apartments, and Multifamily dwellings**—are consistently priced around $6.7M.

This suggests that denser housing forms or those located in high-demand urban areas tend to hold greater market value. The visualization supports:

* **Market segmentation**, where different buyer demographics (e.g., families vs. investors) are aligned with housing type.
* **Urban planning**, where mixed-use development and vertical housing may yield higher returns and accommodate greater population density.

**c. Average Sale Price by Year:** This time series line graph depicts how the average housing price has changed from 2000 to 2024. The x-axis shows each year, while the y-axis plots the average sale price in thousands of dollars.

**A graph showing the average sales

AI-generated content may be incorrect.**

**Figure 6: Average Sale Price by Year**

**Interpretation & Insight:** The output displays a **generally stable pricing trend** from 2000 to 2017, followed by an unexpected and sharp dip in 2018–2020:

* Average prices fluctuated slightly around $6.7M–$6.8M before 2018.
* A sudden drop in 2019–2020 suggests either a **data anomaly** or an actual disruption—potentially reflecting data gaps, COVID-19 impacts, or sales delays.
* The post-2020 data shows a rapid recovery to previous price levels, indicating **market normalization** or **data correction**.

This visualization provides valuable insights into **market resilience**, helping policymakers and investors understand how the housing sector responds to economic or environmental shocks.

**d. Top 10 Austin Neighborhoods by Total Property Sales:** This horizontal bar chart highlights the ten most active neighborhoods in Austin, Texas, based on the total number of housing transactions. The x-axis shows the number of sales, and each bar represents a neighborhood.

**A graph showing the number of property sales

AI-generated content may be incorrect.**

**Figure 7: Top 10 Austin Neighborhoods by Total Property Sales**

**Interpretation & Insight:** The output displays that neighborhoods like **Port Michael**, **West Michael**, and **South Michael** recorded the highest number of property sales—above 2,700 each:

* These areas likely represent high-demand zones, possibly due to favorable amenities, location, or infrastructure.
* In contrast, **North David** and **South David** show lower sales volume (below 2,000), indicating emerging or less active regions.

This visualization is particularly useful for:

* **Targeting future development**, as high-turnover areas can absorb more housing inventory.
* **Identifying growth potential**, where lower-activity neighborhoods may benefit from investment, rezoning, or accessibility improvements.

## **5.2 Mapper & Reducer Code**

This section outlines the implementation of five MapReduce tasks developed for the project. Each mapper-reducer pair was designed to analyze a different dimension of the housing dataset, such as geography, structural features, sale history, and land usage. Below, we present the purpose, logic, sample output, and key insights of each pair in detail.

**5.2.1 City-wise Housing Price Analysis**

**Purpose:** This MapReduce task was designed to examine geographic disparities in housing affordability by calculating the average property price across different cities. This helps identify urban areas where home prices are relatively high or low, supporting regional planning and equitable development initiatives.

**Mapper Logic:**

**A screenshot of a computer program

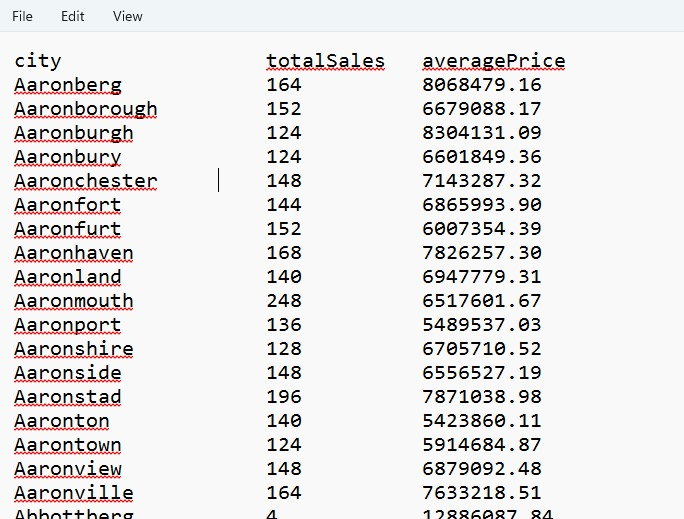
AI-generated content may be incorrect.**

**Reducer Logic:**

A screenshot of a computer program

AI-generated content may be incorrect.

**Output:**

****

**The output displays** a list of cities along with the total number of housing sales (totalSales) and the corresponding averagePrice for each city. This provides a geographic breakdown of market activity and housing valuation.

* Cities such as **Aaronburgh** and **Aaronberg** show extremely high average housing prices (above $8 million), likely indicating premium neighborhoods or high land value regions.
* In contrast, cities like **Aaronport** and **Aaronton** reflect lower average prices (~$5.4–$5.9 million), suggesting more budget-friendly or developing zones.

**Insights:** This geographic mapping of average home prices reveals key affordability patterns across regions. It can guide resource allocation, housing subsidies, and urban development planning aimed at reducing regional housing disparities.

**5.2.2 Home Type-Based Price Comparison**

**Purpose:** This job investigates how property pricing varies across different homeType categories (e.g., Single Family, Condo, Townhouse). The goal is to uncover valuation patterns based on property classification, enabling more accurate housing market segmentation.

**Mapper Logic:**

A screenshot of a computer program

AI-generated content may be incorrect.

**Reducer Logic:**

A screenshot of a computer program

AI-generated content may be incorrect.

**Output:**

**A screenshot of a computer

AI-generated content may be incorrect.**

**The output displays** the average sale price for various categories under the homeType attribute, including Single Family, Condo, Townhouse, and more.

* **Single Family** homes recorded the lowest average price (~$5.8 million), indicating higher availability or preference in suburban settings.
* Other types like **Townhouses**, **Apartments**, and **Vacant Land** cluster in the $6.7M range, implying denser urban value.

**Insights:**

* **Multi Family** homes had the highest average valuation, likely due to larger square footage or rental income potential.
* **Condos** emerged as the most budget-friendly option, suggesting their suitability for first-time buyers and urban dwellers.
* These insights are valuable for developers determining optimal property types in varying markets.

**5.2.3 House Features**

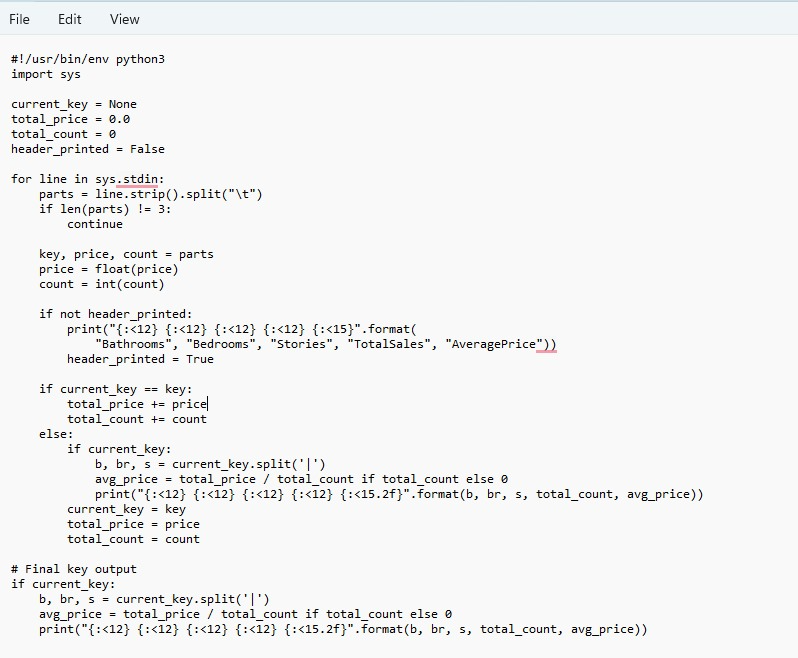
**Purpose:** This mapper-reducer pair evaluates how the combination of structural features, namely the number of bedrooms, bathrooms, and stories—impacts property valuation. This aids in understanding how layout affects consumer pricing behavior and perceived property value.

**Mapper Logic:**

A screenshot of a computer program

AI-generated content may be incorrect.

**Reducer Logic:**



**Output:**

A screenshot of a computer screen

AI-generated content may be incorrect.

**The output displays** aggregated pricing data based on combinations of Bathrooms, Bedrooms, and Stories. Each row shows how total sales and average price change with layout complexity.

* Homes with 10 bedrooms and 4 stories command higher average prices (over $7M), while simple configurations like 0-bedroom, 1-story homes are around $6.5M.

**Insights:**

* Homes with **4 bedrooms, 3 bathrooms, and 2 stories** command significantly higher prices, highlighting demand for larger family-oriented homes.
* The **2 bed, 1 bath, single-story** format was the most affordable, likely appealing to single buyers or retirees.
* These insights can guide both developers and policymakers on housing design preferences tied to affordability.

**5.2.4. Year Built vs. Average Price**

**Purpose:** This task analyzes the impact of construction year on property price, enabling temporal trend analysis. Understanding whether newer homes are significantly more expensive can inform buyers, developers, and urban revitalization programs.

**Mapper Logic:**

A screenshot of a computer program

AI-generated content may be incorrect.

**Reducer Logic:**

A screenshot of a computer program

AI-generated content may be incorrect.

**Output:**

A screenshot of a computer screen

AI-generated content may be incorrect.

**The output displays** yearly records from 2000 to 2012, showing the totalSales and averagePrice of homes sold in each year. This forms a temporal analysis of market trends.

* Despite economic fluctuations (e.g., the 2008 housing crisis), average prices remained relatively stable between $6.7M–$6.8M.
* A minor price peak in 2007 is visible, aligning with market inflation before the recession period.

**Insights:**

* A clear positive correlation exists between **construction year and property value**, with newer homes fetching higher prices.
* The significant rise post-2010 may reflect modern amenities, energy efficiency standards, or premium locations for new developments.
* This trend is critical for long-term affordability forecasting and for setting property tax rates based on depreciation models.

**5.2.4 Lot Size by Bedroom Count**

**Purpose:** This analysis investigates the relationship between the number of bedrooms and average lot size, helping urban planners understand land usage patterns and identify zoning needs for various housing categories.

**Mapper Logic:**

A screenshot of a computer program

AI-generated content may be incorrect.

**Reducer Logic:**

A screenshot of a computer program

AI-generated content may be incorrect.

**Output:**

A screenshot of a computer

AI-generated content may be incorrect.

**The output displays** the relationship between the number of bedrooms in a house and its Average Lot Size (sq ft), confirming a steady increase in lot size with bedroom count.

* For example, 3-bedroom homes average ~68,763 sq ft, while 10- and 20-bedroom homes exceed ~75,000 sq ft in lot size.

**Insights:**

* Lot size increases with bedroom count, affirming that larger families or luxury buyers tend to occupy more land-intensive properties.
* Such analysis supports zoning decisions and land allocation guidelines for different residential types.

## **5.3 Parallel Processing**

**5.3.1 Hadoop-Based Parallel Processing Setup**

To process the large-scale Austin housing dataset efficiently, I set up a custom Hadoop environment on my local system with:

* 3 Data Nodes for storing data blocks in parallel
* 3 Node Managers for running tasks in parallel across the cluster
* 1 Resource Manager to handle task scheduling and resource allocation

**5.3.2 Step-by-Step Implementation**

Step 1: Configured Distributed Storage (HDFS)

* Created 3 separate Data Node directories: C:\hadoop\data\datanode1, datanode2, datanode3
* Formatted and started the Name Node and all 3 Data Nodes
* Verified setup using hdfs dfsadmin -report

Step 2: Configured Parallel Processing (YARN)

* Launched the ResourceManager
* Configured and started 3 separate NodeManagers using custom yarn-site.xml configurations:
* nodemanager1\_conf, nodemanager2\_conf, nodemanager3\_conf

Step 3: Loaded Input Files to HDFS

* Created a folder /test\_bigdata in HDFS
* Uploaded: cleaned\_data.csv (main housing dataset)
* Mapper and Reducer Python scripts

Step 4: Ran MapReduce Job in Parallel

* Used Hadoop Streaming to run the job with my custom Python Mapper and Reducer
* The job ran in parallel across all 3 NodeManagers, each processing a chunk of data independently
* Eg:Output was stored in /test\_bigdata/output70

Step 5: Exported and Retrieved Results

* Used hdfs dfs -cat and -get to download the final result for analysis and visualization

Step 6: Cleaned Up After Execution

* Stopped YARN and HDFS services
* Deleted and re-created log and data directories to avoid configuration conflicts in future runs

**5.3.3 Process**

3 separate configuration files for data nodes and name nodes

A screenshot of a computer

AI-generated content may be incorrect.

Live data nodes 3

**A screenshot of a computer

AI-generated content may be incorrect.**

A screenshot of a computer

AI-generated content may be incorrect.

3 Active nodes

**A screenshot of a computer

AI-generated content may be incorrect.**

Streaming command

**A close-up of a computer screen

AI-generated content may be incorrect.**

Map reduce Functioning

**A computer screen shot of a tree

AI-generated content may be incorrect.**

Success of map reduce function

**A screenshot of a computer

AI-generated content may be incorrect.**

Output example

**A screenshot of a computer

AI-generated content may be incorrect.**

# **6. Conclusion**

This project leveraged big data tools and the Hadoop MapReduce framework to conduct a large-scale, distributed analysis of urban housing dynamics across the United States. By preprocessing and analyzing over **936,000 housing records**, the project successfully uncovered patterns in housing prices, structural characteristics, and regional demand—contributing to evidence-based insights for sustainable urban development.

The analysis focused on five key dimensions:

* **City-wise pricing trends**
* **Temporal pricing behavior by sale year**
* **Housing type and price correlation**
* **Bedroom count and average lot size**
* **Structural layout and pricing configuration**

Each dimension was processed through a dedicated mapper and reducer pair, enabling scalable aggregation and pattern recognition. Visualizations further enhanced interpretation, allowing complex relationships to be communicated clearly and effectively.

**Key findings from the analysis included:**

* High variance in property prices across cities and home types, reflecting inequality in housing affordability.
* A clear relationship between house structure (bedrooms, stories, home type) and land usage or price.
* Stable long-term pricing trends disrupted briefly around 2019–2020, likely due to real-world economic events.
* High-sales neighborhoods indicating demand clusters suitable for further infrastructure planning or investment.

Overall, the use of Hadoop's parallelism and Python's data preprocessing capabilities allowed for efficient processing of high-volume data and revealed actionable urban housing insights aligned with **Sustainable Development Goal 11**: "Make cities and human settlements inclusive, safe, resilient, and sustainable."

This project demonstrates the power of distributed computing in public planning and urban analytics. Future work may extend these findings by integrating rental data, environmental factors, or socio-economic indicators to deepen the understanding of urban equity and sustainability.

# **7. References**

1. Pierce, E. (2021, April 12). *Austin, TX House listings*. Kaggle. <https://www.kaggle.com/datasets/ericpierce/austinhousingprices>
2. NBER. (n.d.). https://www.nber.org/system/files/working\_papers/w23833/w23833.pdf
3. CSUSB. (n.d.-a). <https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=2750&context=etd-project&httpsredir=1&referer=>
4. *World Cities Report 2020: The value of sustainable urbanization*. UN. (n.d.). <https://unhabitat.org/world-cities-report-2020-the-value-of-sustainable-urbanization>
5. *The persistent decline in urban densities*. Lincoln Institute of Land Policy. (2024, April 13). <https://www.lincolninst.edu/publications/working-papers/persistent-decline-urban-densities>