ANALYSIS AND FORECASTING REAL ESTATE SALE PRICES

FINAL PROJECT REPORT

IST 718:- BIG DATA ANALYTICS

M001

GROUP MEMBERS:

KULVEEN KAUR
SUKHAD JOSHI
VAIBHAV GAIKWAD
BISWADIP BHATTACHARYYAA

1. Project Overview

This project analyzes historical real estate transactions and sales data from Connecticut to understand market patterns and forecast future sale prices. Using over two decades of property transaction records (2001–2022), the project applies PySpark, machine learning and time-series forecasting techniques to predict both individual property prices and statewide pricing trends.

The primary goal is to generate insights for stakeholders, investors, realtors, and local governments by analyzing factors such as assessed value, sales ratio, property types, and transaction dates. Our approach integrates both instance-level regression models and aggregate-level time-series forecasting to capture micro and macro trends.

2. Goals

The primary goals of the project are:
Prediction, Inference, and Other Exploratory Goals

1. Prediction Goals:
    - Predict individual property sale prices based on features like assessed value, property type, and sale timing.
    - Forecast the statewide average sale price for the next 24 months using time-series models.
2. Inference Goals:
    - Understand how property features influence sale prices.
    - Investigate regional and temporal variations in the real estate market
3. Exploratory Goals:
    - Generate Visualize trends, seasonality, and anomalies in property sales.
    - Identify deviations from assessed values and notable market shifts.

3. Data Exploration (Including Visualizations)

Dataset Summary:

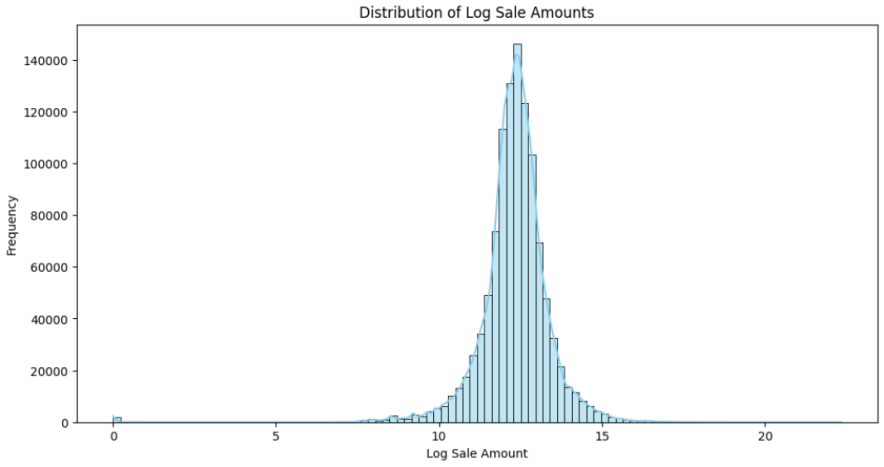
Source: State of Connecticut Open Data Portal (data.ct.gov)

Dataset Name: Real Estate Sales Data
Date Range: January 2001 to December 2022
Total Records: Over 1.2 million property transactions

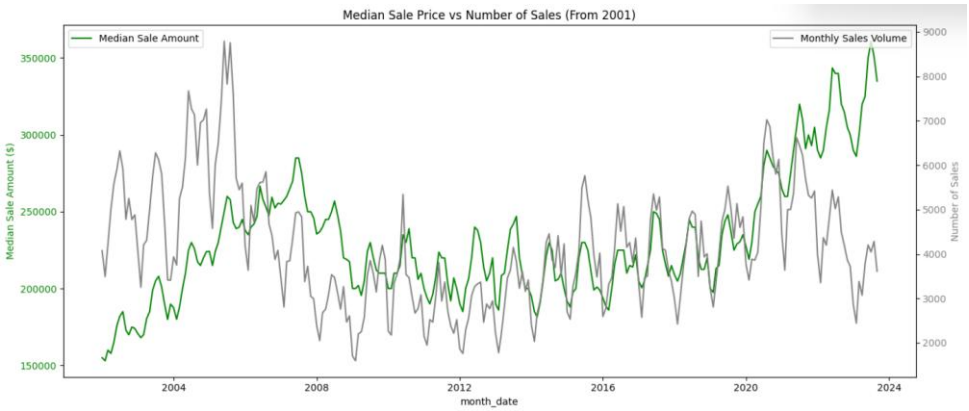
Features:

- Date Recorded: The official recording date of the property transaction, extracted to derive Year, Month, and Quarter for temporal analysis.
- Property Type: Categorical descriptor of property classification (e.g., Residential, Commercial, Industrial).
- Residential Type: Further categorization within Residential properties (e.g., Single Family, Condo, Multi-Family).
- Location (Geo-coordinates): Latitude and longitude points, offering potential for spatial analysis.
- Derived Features:
    - Year, Month, Quarter: Extracted from Date Recorded to enable seasonality and trend modeling.
    - Log Sale Amount: Natural logarithm of Sale Amount to normalize highly skewed price distributions.
    - Log Assessed Value: Log-transformation of Assessed Value for modeling consistency.
    - Log Sales Ratio: Log-transformation of Sales Ratio to manage heavy-tailed behavior.

Visualisations:



- Histogram of log-transformed Sale Amounts and Assessed Values showed much more normal behavior compared to raw values.



- Monthly Median Sale Amount (in green) and Monthly Sales Volume (in gray) for real estate transactions in Connecticut from 2001 to 2023. The graph highlights a steady increase in median prices post-2015, seasonal fluctuations in sales volume, and a noticeable rise in prices following the COVID-19 pandemic.

4. Summary of Methods

The Supervised Regression Modeling

- Tools: PySpark MLlib

- Algorithms: Random Forest Regressor and Gradient Boosted Tree Regressor
- Feature Engineering:
    - Log-transformed Sale Amount and Assessed Value
    - Encoded Property Type and Residential Type using One-Hot Encoding
    - Extracted Year, Month, Quarter from transaction dates
- Performance:
    - Random Forest achieved an R-squared of ~0.92 and RMSE of ~0.24 on log-transformed Sale Amount.

Time-Series Forecasting

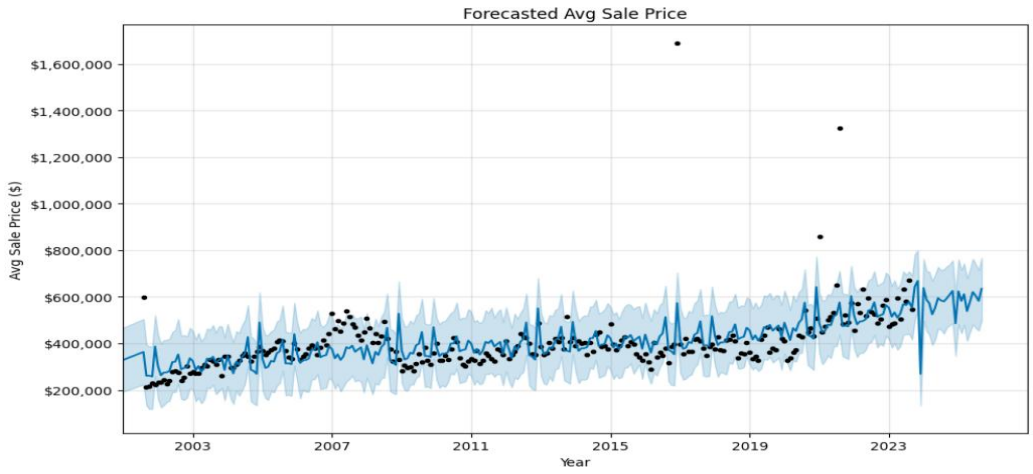
- **Tools:** Facebook Prophet and SARIMA

- **Prophet:** Modeled and forecasted monthly average sale prices through August 2025, capturing long-term trends and seasonality.
- **SARIMA:** Modeled short-term patterns and seasonality for monthly sales volume specifically in commercial properties.
- **Prophet:** Separate Prophet model was used to group the "North Region" subset based on the ‘Lat’ and ‘Lon’ to analyze different Residential Type sales.

5. Results

Prophet Forecast – Statewide Average Sale Prices

Figure1: Forecasted monthly median sale prices for Connecticut real estate using Facebook Prophet, showing a steady upward trend with seasonal fluctuations from historical data through August 2025



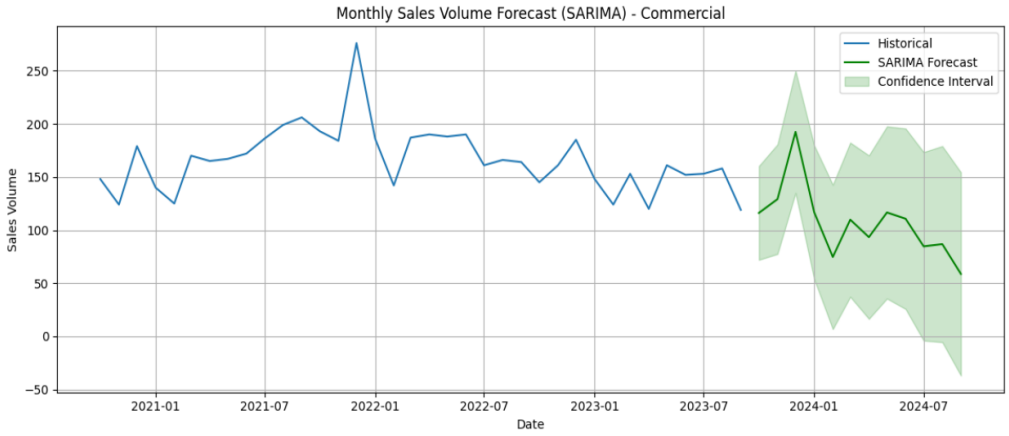
- The statewide Prophet model captured long-term pricing trends and seasonality, predicting a continued rise in property prices over the next two years.

	Date	Predicted Avg Sale Price	Lower Bound	Upper Bound
279	2024-09-30	609289.688108	476916.773475	739016.485842
280	2024-10-31	625334.632817	486574.400044	756445.340316
281	2024-11-30	485840.762802	345380.310344	612784.426456
282	2024-12-31	623366.724360	482664.646525	760688.460387
283	2025-01-31	582124.220170	443410.102948	714961.460043
284	2025-02-28	611084.648328	479755.230924	742774.129436
285	2025-03-31	538543.516962	409786.971107	679073.519630
286	2025-04-30	577313.677998	446548.393817	715614.063583
287	2025-05-31	620473.597490	481279.917829	763002.442359
288	2025-06-30	604260.584535	465741.558276	739567.566623
289	2025-07-31	582104.856781	454325.211006	714795.404747
290	2025-08-31	634406.250197	497665.691861	767687.168298

- This table predicts what average real estate sale prices will likely be for each month from September 2024 to August 2025. It also shows a high and low expected range, giving a realistic idea of how much property prices might vary month-to-monthThis table predicts what average real estate sale prices will likely be for each month from September 2024 to August 2025. It also shows a high and low expected range, giving a realistic idea of how much property prices might vary month-to-month.

SARIMA Forecast – Commercial Property Sales Volume

Figure 2: Forecasted monthly sales volume for commercial property transactions in Connecticut using the SARIMA model, emphasizing seasonal market patterns and short-term fluctuations.



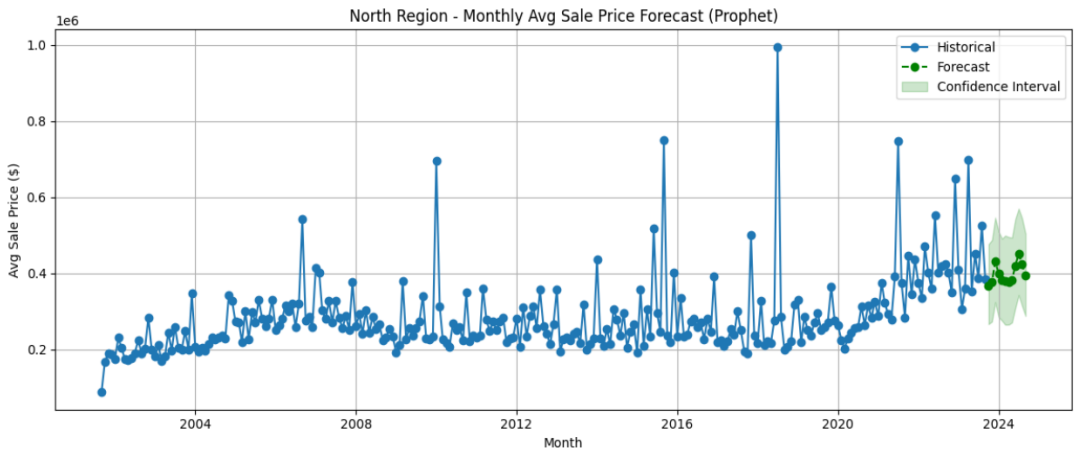
- We used SARIMA on the commercial sales data to spot short-term trends and see how sales change from month to month.

	Date	Predicted Sales Volume	Lower Bound	Upper Bound
0	2023-10-01	116.149717	72.055714	160.243720
1	2023-11-01	129.133246	77.578602	180.687890
2	2023-12-01	192.431361	134.882049	249.980674
3	2024-01-01	116.842015	53.901876	179.782154
4	2024-02-01	74.743204	6.841425	142.644983
5	2024-03-01	109.813339	37.288755	182.337923
6	2024-04-01	93.386218	16.516349	170.256088
7	2024-05-01	116.606066	35.623733	197.588398
8	2024-06-01	110.695735	25.799935	195.591535
9	2024-07-01	84.725835	0.000000	173.362297
10	2024-08-01	86.836610	0.000000	179.059818
11	2024-09-01	58.758307	0.000000	154.404562

-This table predicts how many commercial property sales will likely happen each month from October 2023 to September 2024. It provides an expected range of possible sales, helping businesses and investors prepare for active or slow market periods.

Prophet Forecast – North Region Average Sale Prices

Figure 3: Forecasted median sale prices for the North Region of Connecticut using Facebook Prophet, revealing local price growth trends.



- Regional Prophet modeling uncovered location-specific pricing trends, with the North Region showing sharper seasonal fluctuations but overall price growth consistent with broader patterns.

📍 Forecast Table - North Region

Month	Predicted Avg Sale Price	Lower Bound	Upper Bound	Region
2023-10	366672.354473	266995.901382	476918.732323	North
2023-11	377396.876989	273715.336367	487963.345176	North
2023-12	431234.361372	324364.225609	546470.115703	North
2024-01	400713.074538	285646.525320	509829.786730	North
2024-02	383389.079207	275812.692116	491454.023864	North
2024-03	379018.378954	265422.691900	499562.401812	North
2024-04	378725.501230	265672.115469	495792.446433	North
...				
2024-06	1.027366e+06	651620.382496	1.464717e+06	South
2024-07	9.111002e+05	526865.954942	1.315205e+06	South
2024-08	9.916132e+05	579557.824324	1.398565e+06	South
2024-09	9.024733e+05	486855.889520	1.348859e+06	South

- This table predicts the average monthly sale prices specifically for the North Region of Connecticut from October 2023 onward. It provides both the predicted values and a confidence range for future prices, helping capture local market dynamics more precisely than statewide models.

6. Problems Encountered

We encountered the following challenges during our project:

1. Large Dataset Handling:
   The full real estate dataset contained over 1.2 million records, which caused memory and performance issues when working in Google Colab. To solve this, we used PySpark, which allowed the data to be processed in a more distributed and efficient way.

2. Date Formatting Challenges:
   Some records in the dataset had inconsistent or missing date fields, which made it difficult to prepare the data for time-series models like Prophet.

3. Feature Vector Assembly:
   Building a clean feature set for machine learning models required combining different types of data including numerical values and one-hot encoded categories. This involved multiple transformation steps in PySpark to create a single feature vector that models like Random Forest and Gradient Boosted Trees could use.

4. Outlier Sensitivity:
   The dataset contained a few extremely high-priced luxury sales, which heavily skewed the average sale price values. While log transformations helped to reduce the impact of these outliers, their influence could still be seen in some raw data trends and visualizations.

7. Discussion and Conclusion

In this project, we explored and predicted real estate sale prices in Connecticut using a combination of machine learning and time-series forecasting methods. By analyzing over 1.2 million property transactions, we were able to uncover important patterns, such as the steady rise in median sale prices over the past two decades and the seasonal trends in the number of sales. Random Forest and Gradient Boosted Tree models gave strong predictive performance at the property level, while the Prophet and SARIMA models helped forecast broader trends across time. Throughout the process, we faced challenges with large dataset handling, date inconsistencies, and outlier sales, but careful preprocessing and transformation steps helped address these issues. Overall, combining both property-specific predictions and market-wide forecasting gave a more complete understanding of how Connecticut's real estate market behaves and evolves.

This project successfully met its goals of both predicting individual property sale prices and forecasting future market trends. Using PySpark and advanced modeling techniques, we handled large datasets efficiently and achieved strong prediction results, with clear seasonal patterns and growth trends reflected in the forecasts. The Prophet model's future projections suggest that property prices will continue to rise steadily into 2025, supporting current market observations. Our analysis shows that with the right combination of preprocessing, feature engineering, and model selection, even large and complex real estate datasets can be used to gain valuable business and market insights. Future improvements could include adding external economic factors like interest rates to make predictions even stronger.

8. Citations

Connecticut Office of Policy and Management. "Real Estate Sales Data (2001–2022)." *Connecticut Open Data Portal*, 2024, https://catalog.data.gov/dataset/real-estate-sales.

Taylor, Sean J., and Benjamin Letham. "Forecasting at Scale." *PeerJ Preprints*, vol. 5, 2017, doi:10.7287/peerj.preprints.3190v2.

Apache Software Foundation. "Apache Spark™ – Unified Analytics Engine for Big Data." *Apache Spark*, 2024, https://spark.apache.org/.

Meta Platforms, Inc. "Facebook Prophet: Forecasting at Scale." *Meta Open Source*, 2024, https://facebook.github.io/prophet/