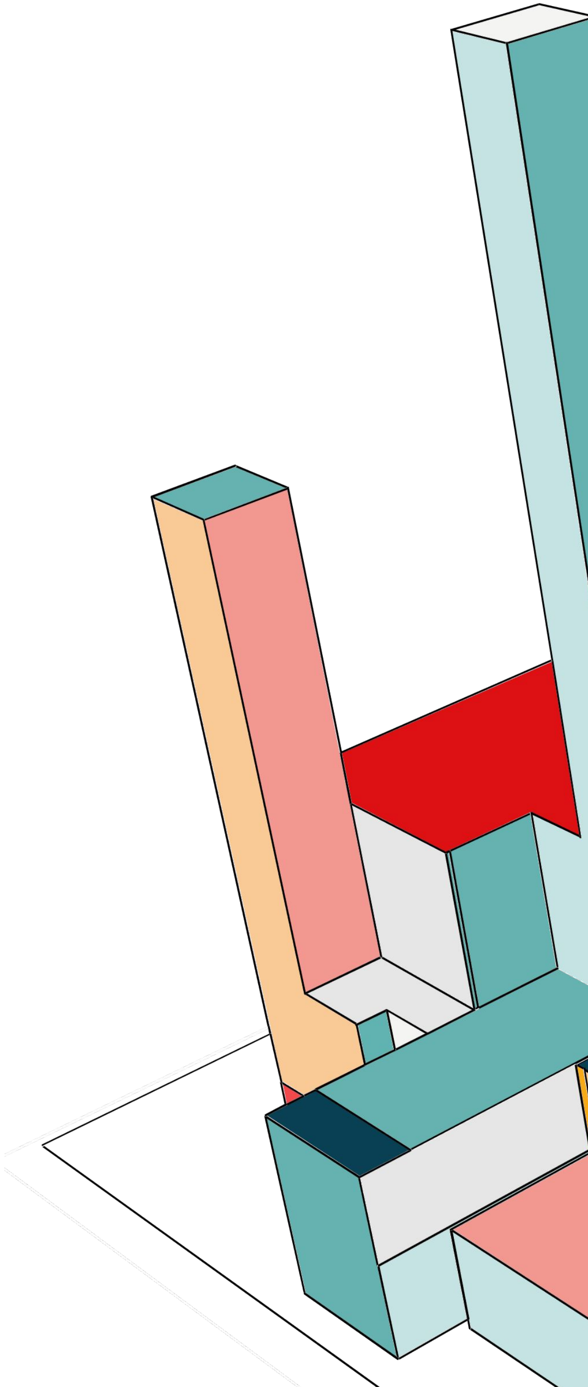
RedfinPredict

Contributors:
Rome Lin, Hongfan Lu, Aishwary Jadhav, Maitreyi Ekbote, Ching-Ping Chan

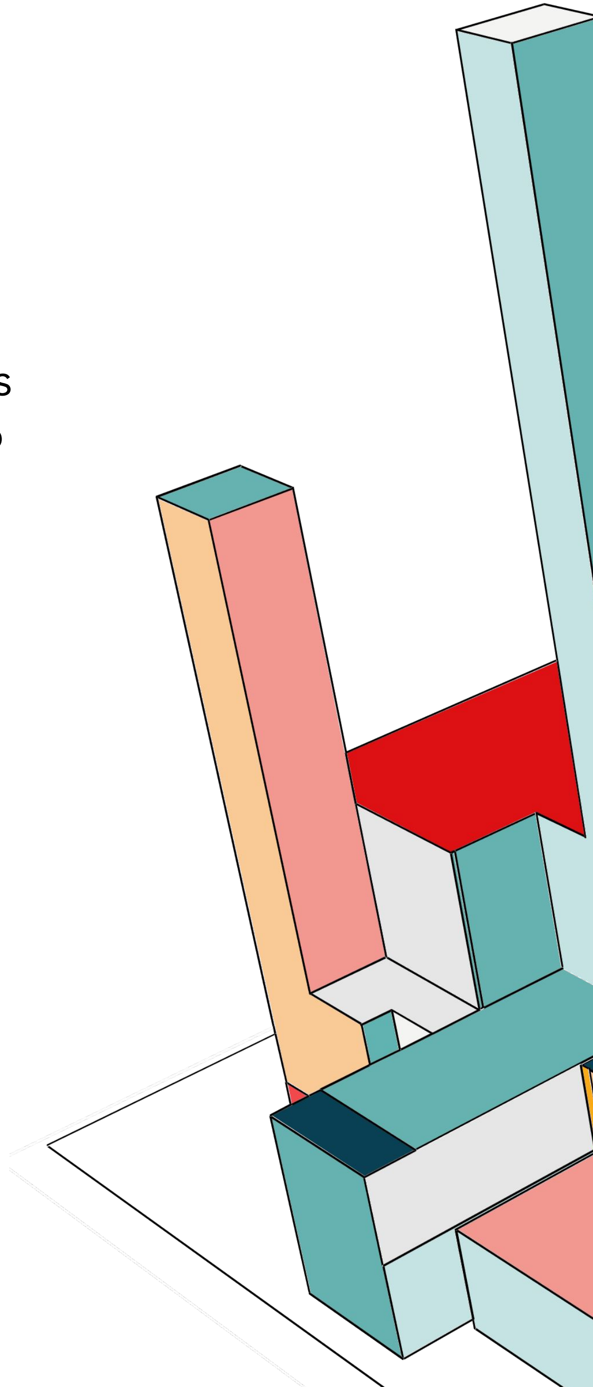
Table of Content

1. Problem statement description
2. Human centered description of users or use cases
3. Technologies you used
4. Major challenges
5. Next steps



PROJECT DESCRIPTION

- GOAL:
    - Build a tool that offers real-time predictions of the median sale prices for homes in popular U.S. cities, based on data from Redfin's housing market from 2019 to the present
- OBJECTIVES:
    - Track Historical and Current Price Trends (2019-Present)
    - Simplify Data Interpretation
        - Present data in a visual format and enable interactions for users to easily explore trends, compare predictions, and make informed decisions
    - Provide Accurate Price Predictions
        - Dashboard allows users to select the method that best reflects their perception of market trends



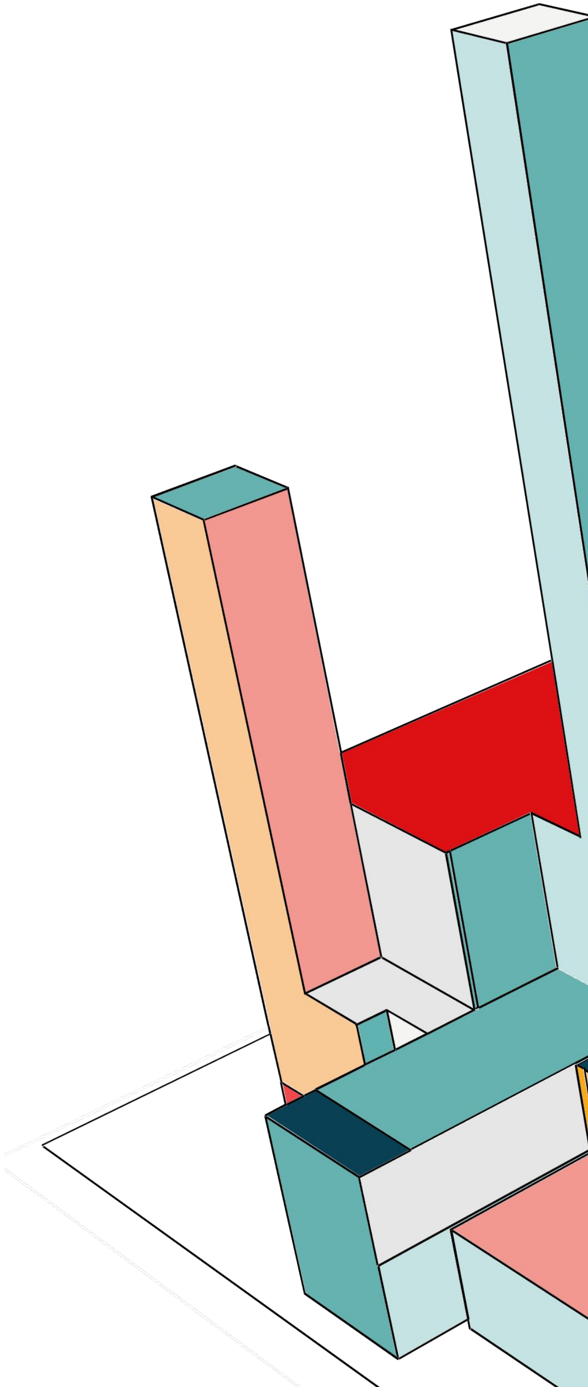
USE CASES

Target Users & Goals

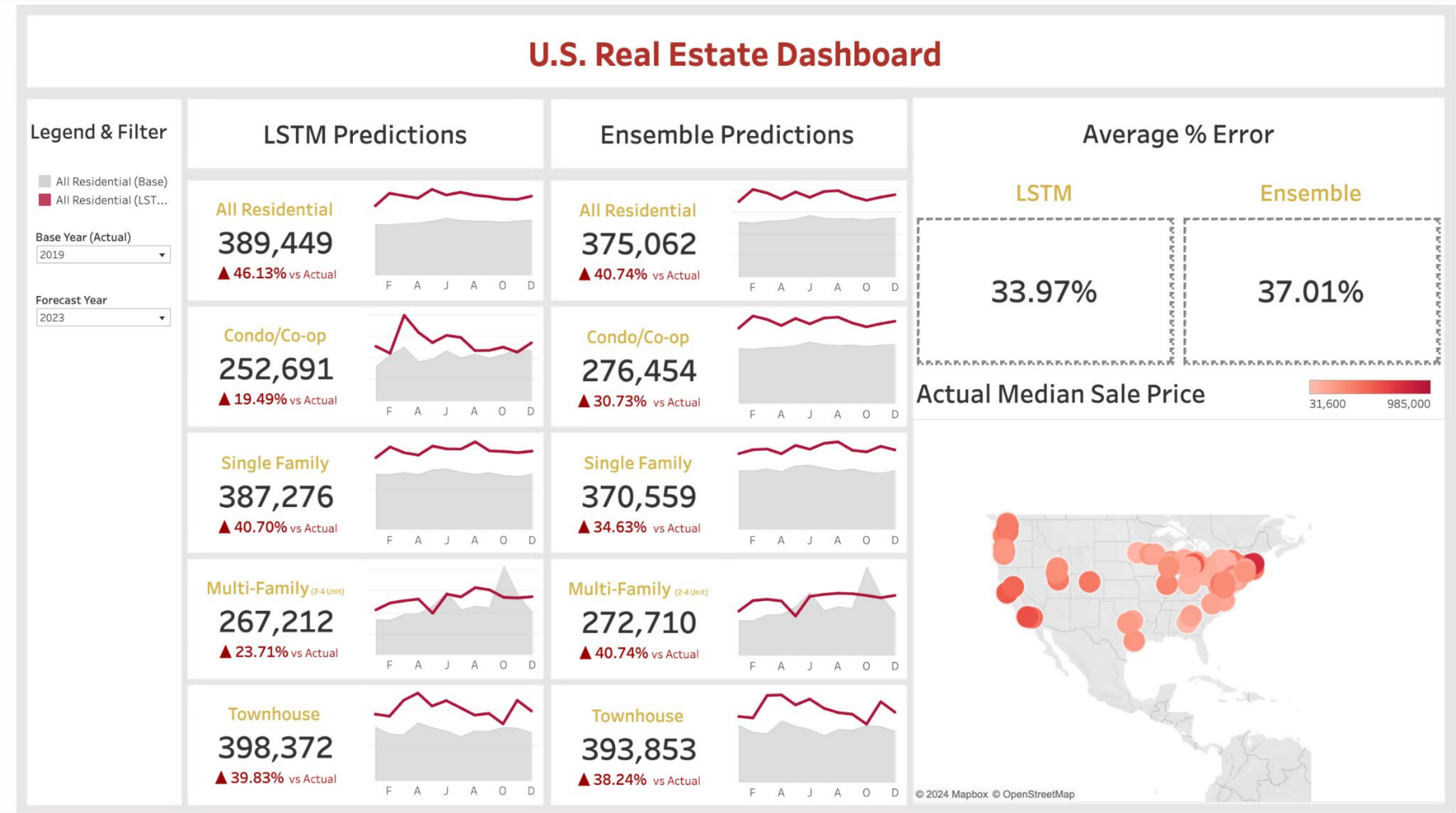
- Homebuyers
  - Find affordable properties
- Investors
  - Maximize on return on investment
- Real-Estate Agents
  - Provide data-driven advice to clients

Use Cases

- Visualize Historical Trends
- Leverage machine learning predictions to guide purchasing decisions



DASHBOARD



U.S. Real Estate Dashboard

Legend & Filter

▨ All Residential (Base)
🟥 All Residential (LST...

Base Year (Actual): 2019

Forecast Year: 2023

	LSTM Predictions	Ensemble Predictions
All Residential	389,449 ▲ 46.13% vs Actual	375,062 ▲ 40.74% vs Actual
Condo/Co-op	252,691 ▲ 19.49% vs Actual	276,454 ▲ 30.73% vs Actual
Single Family	387,276 ▲ 40.70% vs Actual	370,559 ▲ 34.63% vs Actual
Multi-Family (2-4 Unit)	267,212 ▲ 23.71% vs Actual	272,710 ▲ 40.74% vs Actual
Townhouse	398,372 ▲ 39.83% vs Actual	393,853 ▲ 38.24% vs Actual

Average % Error

LSTM	Ensemble
33.97%	37.01%

Actual Median Sale Price

© 2024 Mapbox © OpenStreetMap

https://bit.ly/cse583project

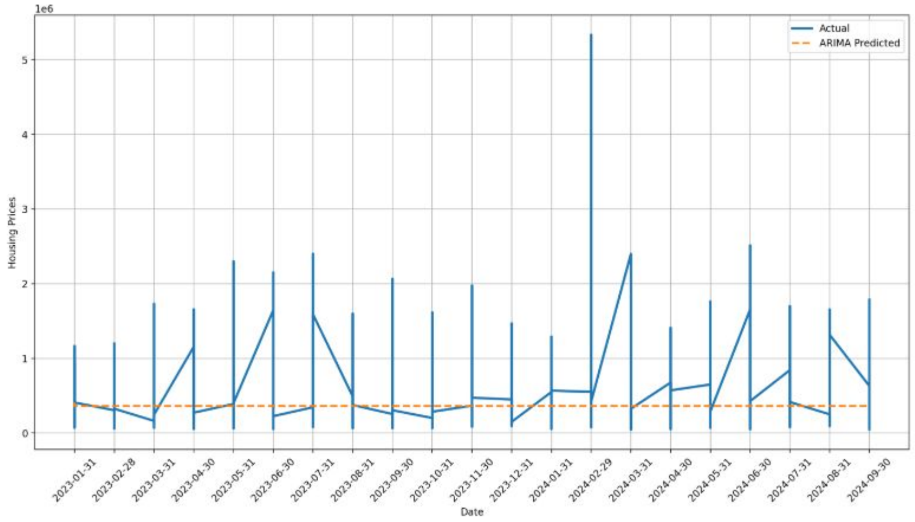
Arima - Sarimax Models

Basic Time-series Prediction

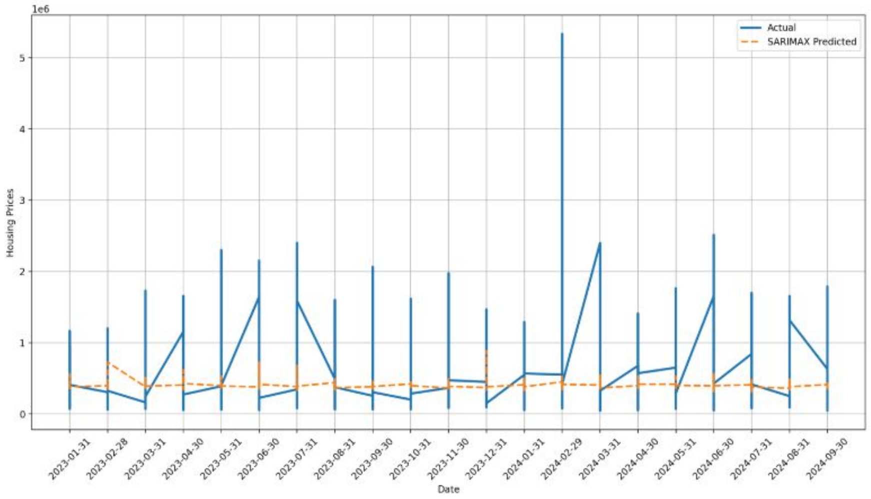
ARIMA & SARIMAX

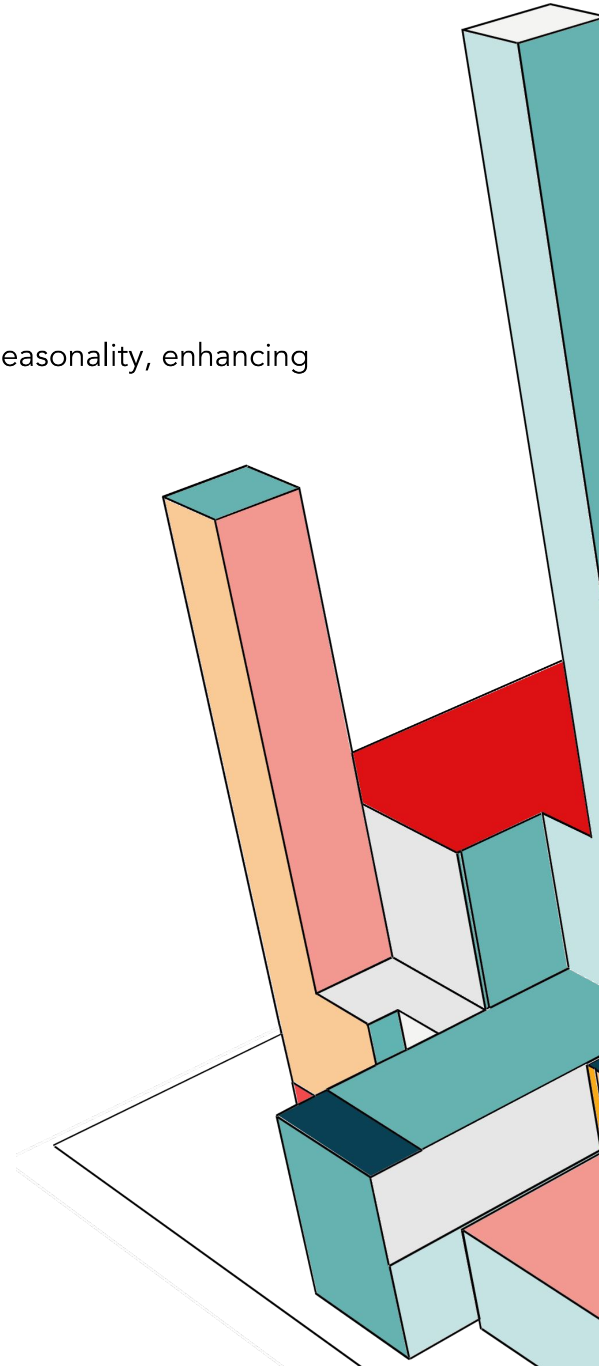
- Technologies you used:
  - ARIMA: A univariate time series model using only the target variable. (p, d, q)
  - SARIMAX: An extension that incorporates exogenous variables and seasonal patterns (P, D, Q, m).
  - Tools: pmdarima for automated order selection, statsmodels for training, and sklearn for metrics.
  - Progression involved expanding ARIMA's capabilities to SARIMAX by integrating explanatory variables and addressing seasonality, enhancing predictive power.
- Major challenges
  - Seasonality Detection: Balancing seasonal and trend components for SARIMAX models.
  - Exogenous Variable Selection: Identifying meaningful variables for SARIMAX.
- Next steps:
  - Hyperparameter Optimization: Refine SARIMAX parameters further using grid search.

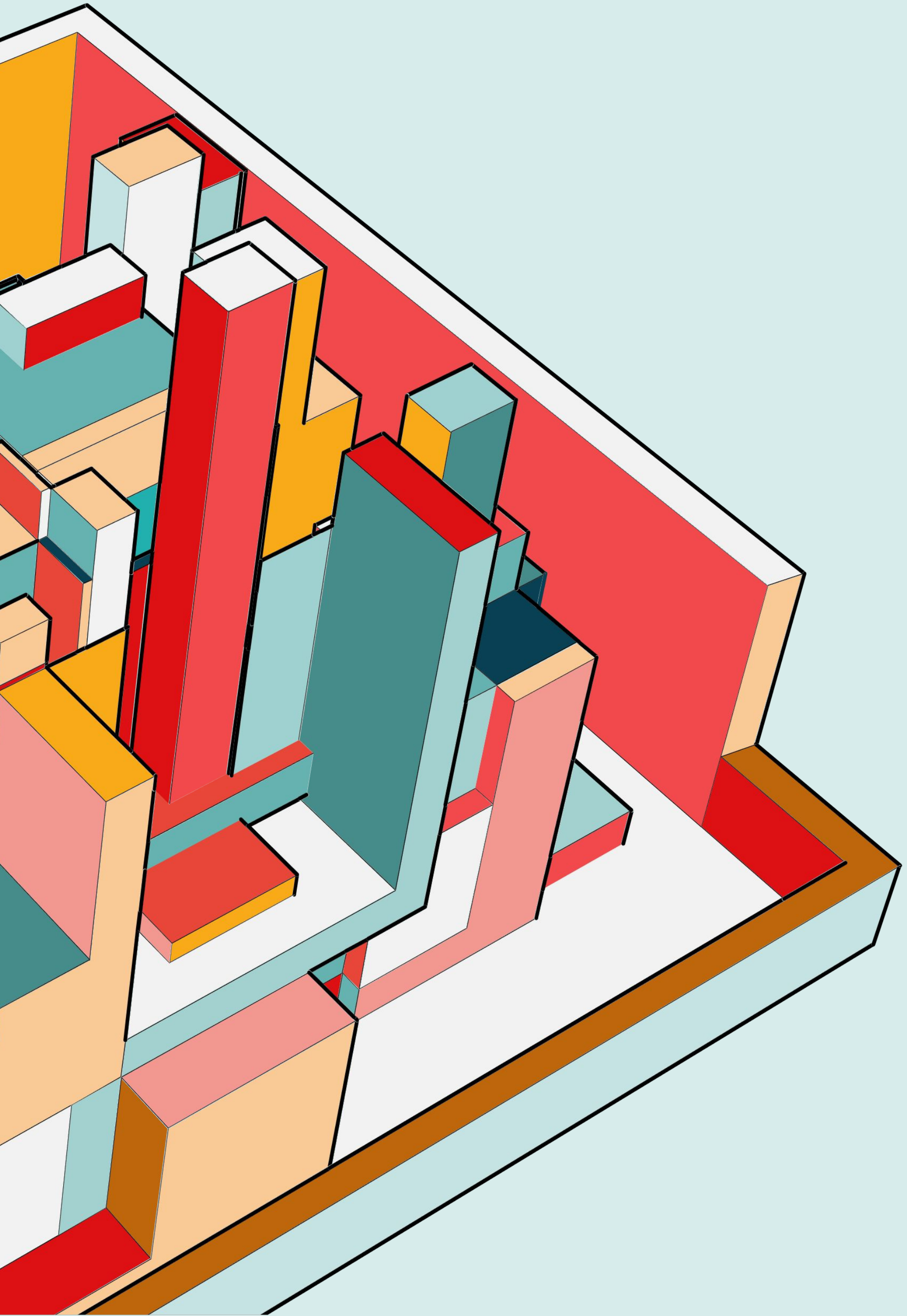
ARIMA Predictions vs Actual



SARIMAX Predictions vs Actual







Random Forests

Time Series Forecasting

Using Random Forest, we deliver actionable insights to optimize pricing strategies and enhance profitability.

Why Random Forest?

- Captures complex patterns in data, ideal for dynamic price forecasting.
- Identifies key drivers of price trends for actionable business insights.
- Adaptable to time series tasks, overcoming limitations of traditional models.

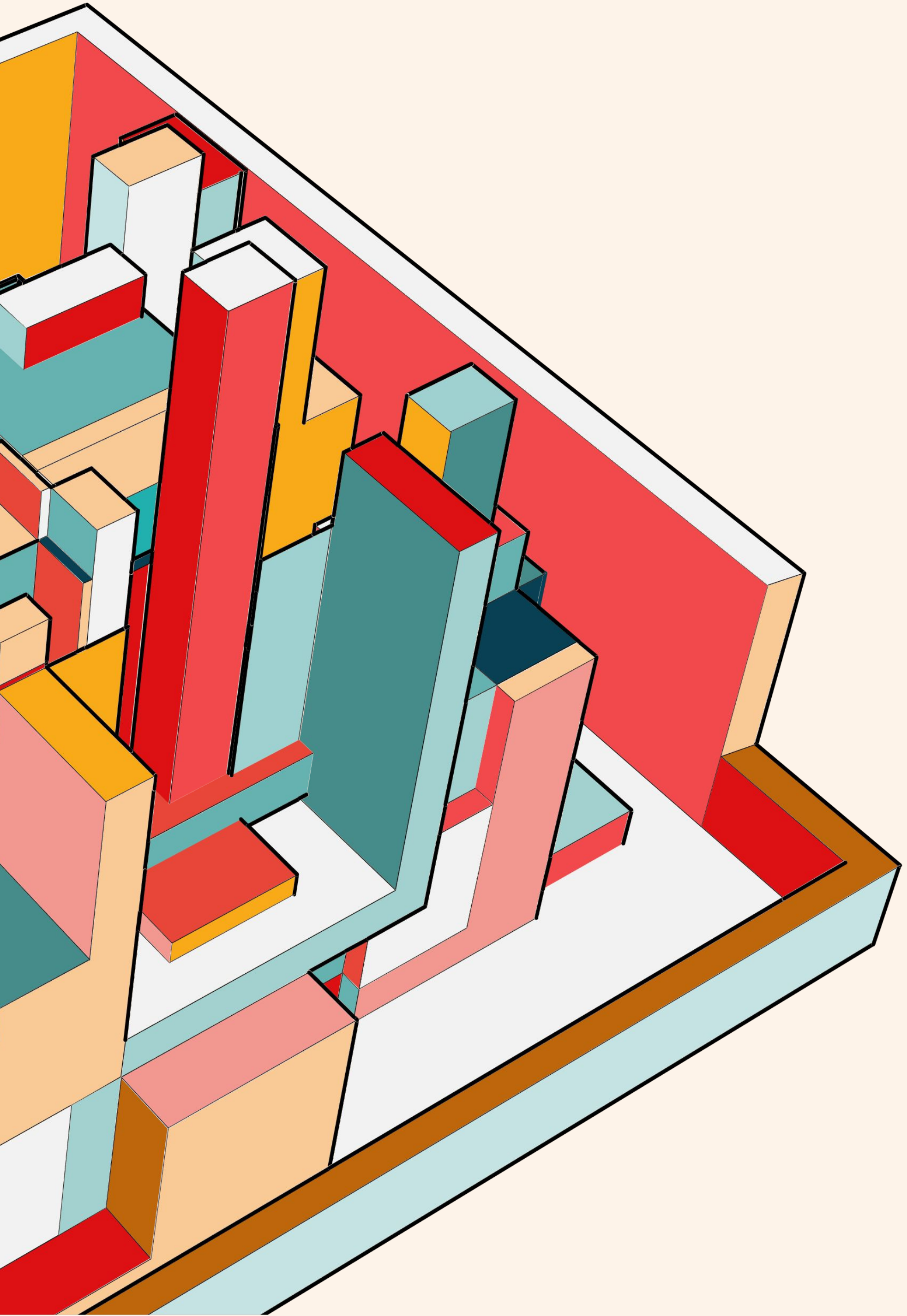
Our Approach

- Focused on predictors like market heat index, sale price trends, and inventory.
- Tuned hyperparameters n_estimators, max_depth, max_features, min_samples_split, and min_samples_leaf to optimize model accuracy and generalization. using GridSearch to ensure best performance.
- Assessed model reliability through robust metrics like R square and MAPE for forecasting accuracy.

Key Insights

Achieved high predictive accuracy and identified key drivers like list price

Model supports strategic decisions, and scales for multi-city forecasting and planning.



Long Short-Term Memory (LSTM)

Deep Learning : RNN

Time-Series Forecasting

LSTM

Why LSTM?

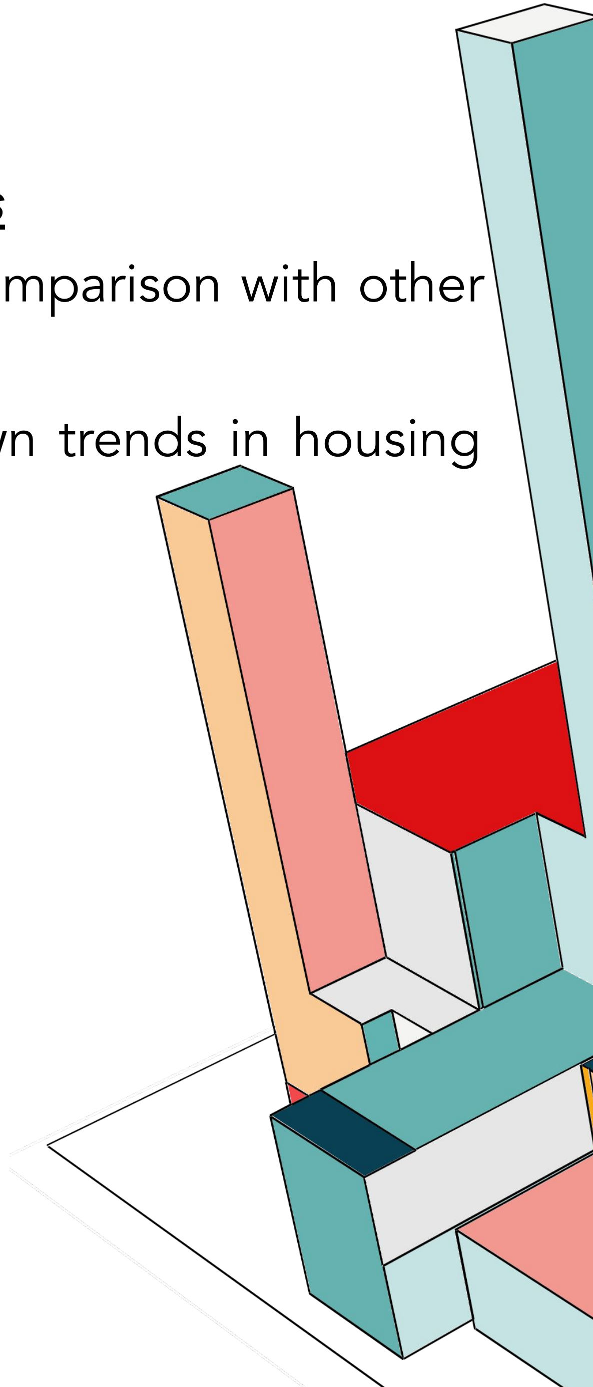
- Captures non-linear complexity patterns.
- Ideal for dynamic, multi-variable datasets

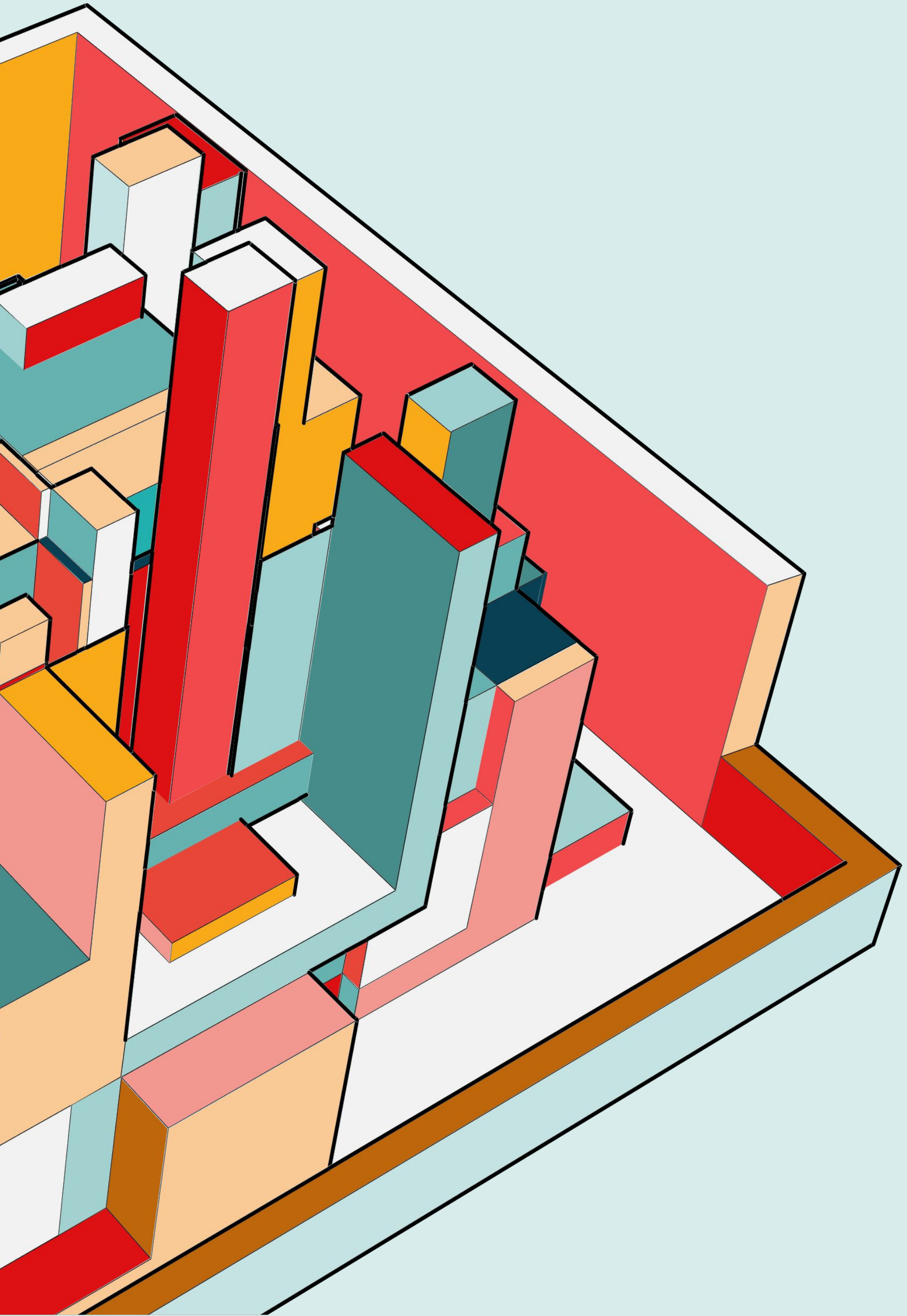
Our Approach

- Scaled input using MinMaxScaler.
- Train with LSTM layers, dropout regularization, and tuned hyperparameters (units, dropout rates, and batch size)
- Assessed performance through <u>RMSE</u> and <u>MAPE</u>.

<u>Key Insights</u>

- Best performing model in comparison with other models
- Able to capture up and down trends in housing prices





Challenges & Future Work

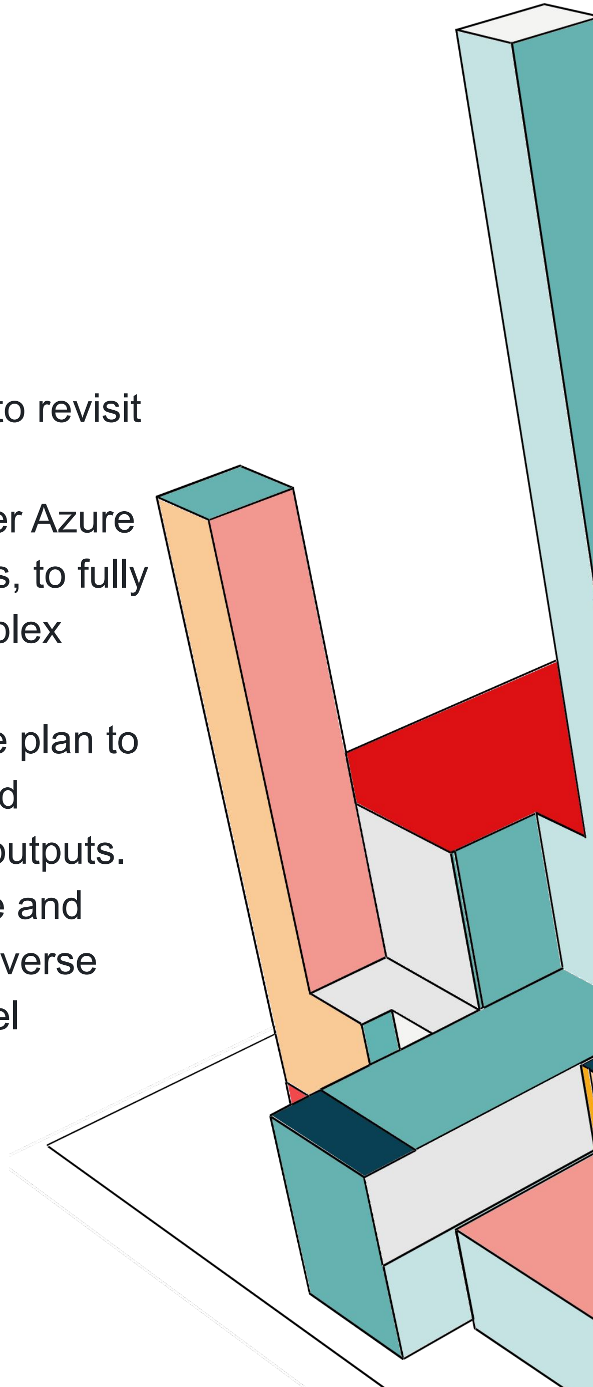
Challenges & Future Work

Challenges:

- **Cloud Deployment Limitations**: We initially planned to deploy the project on Azure Cloud to have more computing resources for more complex model, but we faced resource limitations, particularly with the free-tier services.
- **Machine Learning Approach**: Currently, our machine learning models' outputs require manual adjustments to consolidate the predictions into a single file, final_ML_prediction_for_plotting.csv, for plotting and visualization. This is mainly due to inconsistencies in the output formats across different models done by different team members.

Future Work:

- **Improved Cloud Deployment:** We plan to revisit cloud deployment options with increased resources, either by upgrading the free-tier Azure account or exploring other cloud platforms, to fully leverage scalable compute for more complex models.
- **Machine Learning Synchronization**: We plan to align on output formats across models and automate the consolidation of prediction outputs. As for prediction accuracy, we will explore and test additional models and gather more diverse data for training, as our current best model (LSTM) still has over 30% error.



Thank you!