Milestone 3: Evaluation, Interpretation, Tool Development, and Presentation

1. Objective

The real estate market is a dynamic and multifaceted domain shaped by various economic, geographic, and social factors. In this project, I aim to analyze and predict real estate trends using historical housing data, enriched with visualization and modeling techniques. The goal is to uncover meaningful patterns in property prices, identify the key drivers of value, and provide actionable insights for buyers, sellers, and policymakers.

Real estate prices are influenced by features such as location, property size, number of rooms, year built, and nearby amenities. Through thorough preprocessing, exploratory data analysis, and machine learning models, this project evaluates how these attributes correlate with housing prices. By building interactive dashboard, I strive to make complex real estate data more accessible and useful for decision-making.

Github link -> https://github.com/SaiPande/cap5771sp25-project

2. Dataset Description

Datasets used are->
Kaggle Dataset ->

Real Estate Sales 2001-2021 https://www.kaggle.com/datasets/utkarshx27/real-estate-sales-2001-2021-gl (License : CC0: Public Domain)

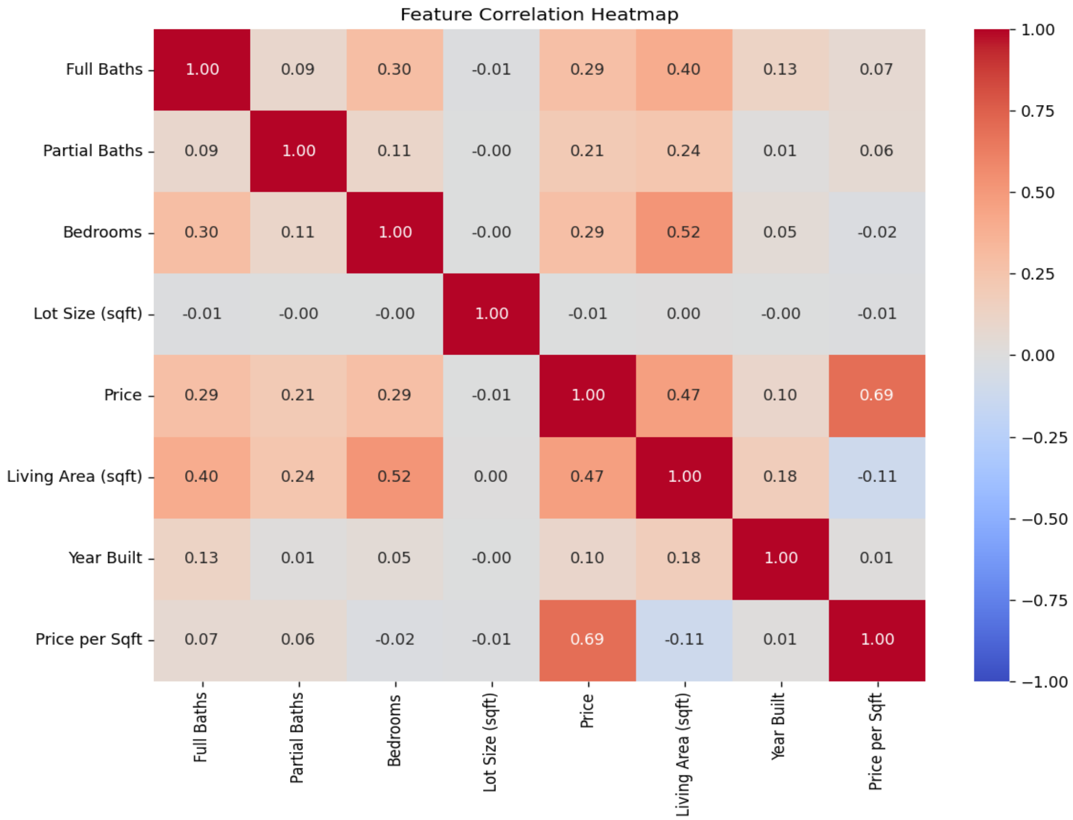
Zillow Economics Data https://www.kaggle.com/datasets/zillow/zecon (License : Data files © Original Authors)

Real Estate DataSet **https://www.kaggle.com/datasets/arslanali4343/real-estate-dataset** (License : CC0: Public Domain)

3. Recap of Milestone 1 and 2

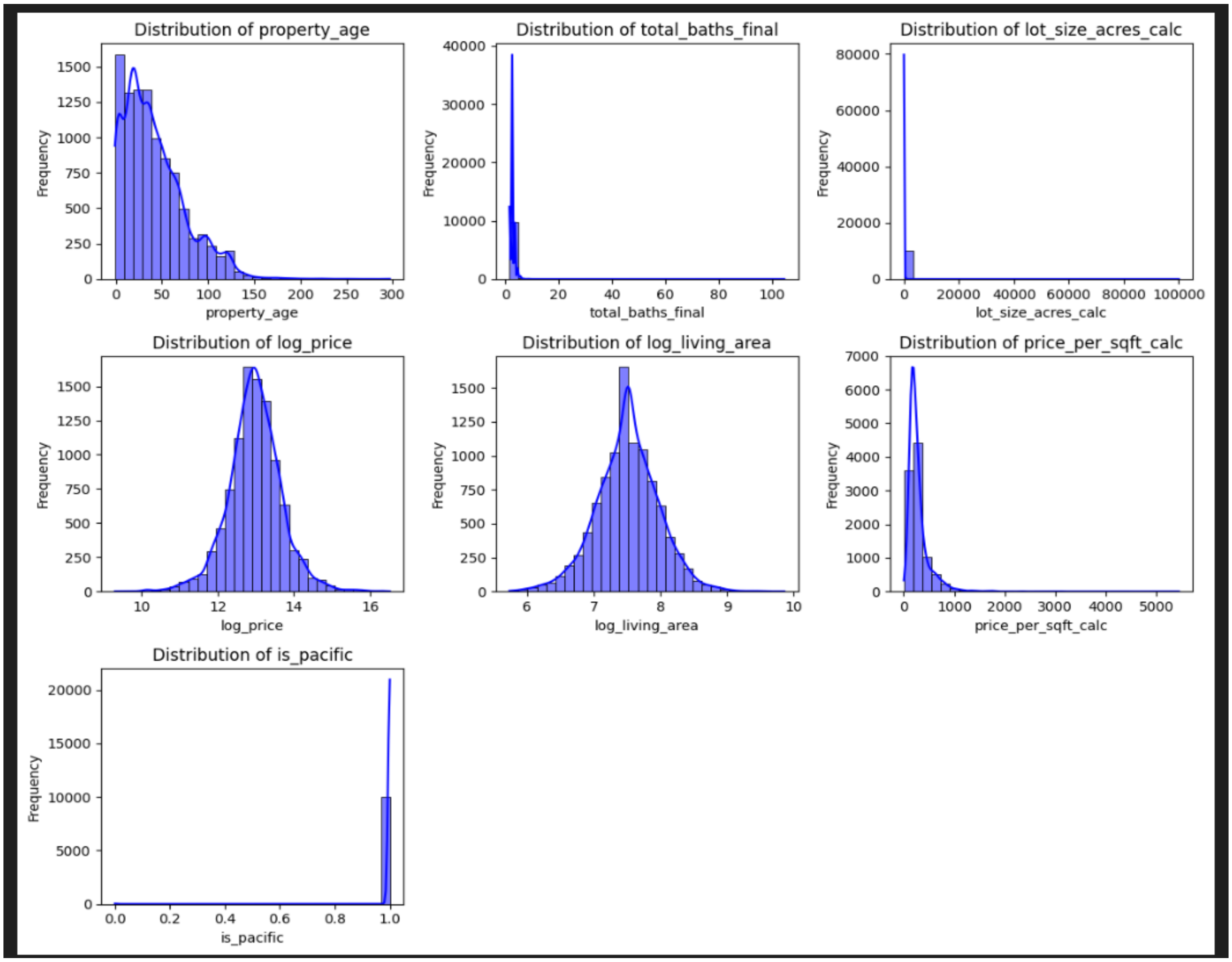
Milestone 1: Data Cleaning, Preprocessing, and Exploratory Data Analysis

In Milestone 1, the primary focus was on preparing the real estate dataset for modeling by performing comprehensive data cleaning, preprocessing, and exploratory data analysis (EDA). Data cleaning involved addressing missing values by filling missing values in critical columns such as beds, baths, and living_area_sqft, and removing rows with missing target values like price. I also corrected data types by converting fields like snapshot_date and data_gen_date into datetime formats, ensuring proper time-based analysis later. To handle outliers, I applied capping strategies, such as limiting the price variable to the 99th percentile to remove extreme high-value distortions that could skew the model. Preprocessing steps included creating new variables like price_per_sqft to standardize property pricing across different home sizes, and label encoding categorical variables such as property_type and state for compatibility with machine learning models. Once the data was clean and consistent, I conducted a detailed EDA to better understand the dataset's structure. This involved visualizing distributions of key variables like price, lot size, and living area, and exploring trends and patterns across states, cities, and property types. Through these steps, I identified important relationships and prepared a robust dataset for feature engineering and modeling.



Milestone 2: Feature Engineering, Feature Selection, and Modeling

In Milestone 2, the focus shifted toward enhancing the predictive capabilities of the dataset through feature engineering, feature selection, and model development. I created new features to capture more meaningful real estate patterns, including property_age (difference between the snapshot year and year built), price_per_sqft (price divided by living area), and is_pacific (binary indicator for properties located in the Pacific time zone). Log transformations were also applied to skewed variables like price and living_area_sqft to normalize their distributions and improve model stability. For feature selection, I prioritized a subset of informative predictors such as beds, property_age, total_baths_final, living_area_sqft, lot_size_acres_calc, price_per_sqft_calc, and is_pacific to balance model performance and generalization. With the finalized feature set, I trained three different models — Linear Regression, Random Forest Regressor, and XGBoost Regressor — to predict property prices. Each model was evaluated based on key metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R² Score, and ROC AUC to ensure a thorough performance comparison. This milestone significantly improved model accuracy, highlighted the strengths of ensemble methods over simple regression, and set up the foundations for building an interactive predictive dashboard.



4. Model Evaluation Summary (Test Set)

Metric	Linear Regression	Random Forest	XGBoost
MAE	120,717.90	8,963.72	21,954.10
RMSE	237,612.46	70,740.08	177,695.31
R² Score	0.7957	0.9819	0.8858
ROC AUC	0.9643	0.9999	0.9995

5. Interpretation Summary

- Random Forest is clearly the best performer with the lowest error (MAE & RMSE) and highest R^2. It captures patterns extremely well.
- XGBoost is also strong, with better generalization than linear regression and fairly low prediction errors. It balances power and flexibility well.
- Linear Regression is a good baseline but underperforms in comparison. It misses complex relationships, leading to larger prediction errors.

MAE vs RMSE Trade-Off:

- MAE (Mean Absolute Error) represents the average magnitude of errors in predictions, treating all errors equally. It is robust to outliers.
- RMSE (Root Mean Squared Error), on the other hand, penalizes larger errors more heavily due to squaring, which helps detect whether the model is making a few severe mistakes.
- For Linear Regression, we observe a significant gap between MAE (≈ $120,718) and RMSE (≈ $237,612). This wide gap indicates that while most predictions are close to actual values, the model struggles with large outlier errors, leading to higher penalty under RMSE.
- In contrast, the Random Forest model has an MAE of only ≈ $8,964 and an RMSE of ≈ $70,740 — much closer together, indicating that errors are more evenly distributed and the model is better at avoiding severe mispredictions.

- XGBoost, with an MAE of ≈ \$21,954 and RMSE of ≈ \$177,695, strikes a middle ground — showing stronger performance than linear regression, but still somewhat sensitive to high-value errors (e.g., luxury property pricing).

R^2 Score Interpretation:

- The R^2 score measures the proportion of variance in the target variable (price) that is explained by the model's inputs.
- The Random Forest achieves an R^2 of 0.9819, the highest among all models, indicating it captures almost all variability in property prices.
  XGBoost also performs well with an R^2 of 0.8858, showing strong predictive power with potentially better generalization due to regularization.
- Linear Regression, with an R^2 of 0.7957, performs adequately but falls short compared to ensemble models — suggesting it is not able to capture the complex nonlinear relationships in the data.

Residual Analysis:

To better understand the distribution of prediction errors, I visualized the residuals (actual price minus predicted price) against the predicted prices using a scatter plot (see Figure below). The plot shows that most residuals are concentrated around zero for lower and mid-range predicted prices. However, for very high predicted values (e.g., luxury properties above \$5 million), residuals become more spread out — with both underpredictions and overpredictions exceeding \$1 million in some cases.

This pattern tells us that the model performs consistently on average listings but exhibits mild heteroscedasticity — i.e., the variance of prediction errors increases with the predicted price. This is a common phenomenon in real estate datasets due to the broader price range and unique characteristics of luxury properties. In future work, I will try applying log-transformations or training separate models for high-value segments could mitigate this issue.

Model Complexity vs Interpretability

- Random Forest & XGBoost are black-box models. While accurate, they're harder to interpret.

- Linear Regression is fully transparent — coefficients directly show how each feature affects price. This makes it better when we need explainability (e.g., in policy settings or lending).

Bias-Variance Analysis for Random Forest and XGBoost

From a bias-variance tradeoff perspective, Random Forest tends to have **low bias and moderate variance**, meaning it fits training data very well but can slightly overfit on noisy signals. XGBoost offers a better **bias-variance balance** by introducing regularization during boosting iterations, resulting in slightly more bias but reduced variance. This explains why Random Forest achieves the lowest errors here, but XGBoost may generalize slightly better to unseen datasets.

Model Selection Considerations

- Deployment time: RF and XGBoost require more memory and are slower to predict on large datasets.
- Data size: XGBoost scales well with big data; RF is better when feature count is high.

The Future Work on my Project would be:.

- Using feature importance or SHAP values to explain decisions from RF/XGB.
- Evaluate residual plots vs predictors to detect any unmodeled non-linearity or heteroscedasticity.
- Adding more datasets like weather, crime rate, nearby schools and office, transportation services to find best real estate price deals and properties based on these all factors which relates to the real world.

Tool: Interactive Real Estate Dashboard using Streamlit- Purpose and Justification

Purpose

The Streamlit dashboard serves as the central tool for presenting all analysis, modeling, and prediction outputs in a structured and interactive format. It enables users (e.g., analysts, students, instructors) to:

- Explore the dataset visually without needing technical expertise.
- Assess model performance at both macro and micro levels.
- Interact with predictions based on real property attributes.

The dashboard improves transparency, usability, and engagement, making model insights accessible to a broader audience.

What does my Interactive Dashboard Shows and Why

1. Sidebar: State Filter

Why: Real estate dynamics vary significantly by state. This filter lets users isolate and analyze market trends within a specific geographic region, ensuring localized insights rather than diluted national averages. The user can select “All States” or states of their choice (like FL for Florida) to see the real-estate data and prediction.

2. The Sliders for users to see the property prices based on the Growth Rate and the Years into future:

Growth Rate (%) slider:
This lets the user set an annual price growth rate (like 4% per year). The users can simulate how much property prices will rise each year into the future based on this percentage.

Years into Future slider:
This lets the user pick how many years ahead they want to forecast the property price (e.g., 10 years into the future).

2. Key Performance Indicators (Top-Level Metrics)

- Average Price: Represents the central tendency of property values in the selected state.
- Median Lot Size (Acres): Reflects property scale, often tied to rural vs urban differentiation.
- Average Days on Market: Indicates market fluidity. Longer durations imply lower demand or overpricing.

Why: These KPIs quickly summarize current market conditions and provide a high-level health check of the dataset.

3. Price Distribution (Histogram)

Why: Highlights the spread and skewness of property prices. Helps identify whether the market is dominated by affordable or luxury homes and whether log transformation was appropriate for modeling.

4. Living Area vs Price (Scatter Plot)

Why: Captures the correlation between size and price. Useful for validating assumptions like "larger homes cost more" and spotting pricing outliers that may need further review or exclusion. There is an awesome feature where the user can select or deselect whichever cities they want and see the plot with only the required cities.

Why: Well, if the user has decided or planned to live in 3-4 cities and not considering other cities to live, then it makes sense to only compare these 3-4 cities of a state and see the picture with more clarity.

Prediction Section – Interpretation and Utility

5. Predicted vs Actual Price

Why: Evaluates how close model predictions are to real prices. Identifies systemic under- or over-prediction trends and helps visually assess accuracy.

6. Residual Distribution

Why: Shows the distribution of prediction errors. Reveals if the model is consistently biased in one direction or if errors are centered and well-behaved.

Error Heatmap by Price and Size

Why: Allows 2D analysis of error distribution. Identifies combinations (e.g., large homes at low prices) where the model underperforms.

7. Predicted Price Map

Why: Plots predicted prices geographically. Allows validation of model output against expected regional price levels and identifies geographic prediction discrepancies.

8. Average Price by City (Bar Chart)

Why: Ranks cities to highlight regional price disparities. Supports comparative analysis for investment decisions or regional targeting strategies.

9. Property Type Distribution (Pie Chart)

Why: Gives an overview of inventory composition. Knowing whether the dataset leans toward single-family homes, condos, or townhomes helps frame model expectations and segment-specific performance. This also helps user to see what kind of properties are commonly available to make a decision or to know how difficult it would be to find a property which is not of a common type like townships. (in our data, single-family homes in most of the states)

10. Days on Market by City (Boxplot)

Why: Shows market responsiveness in different cities. Useful to detect lagging markets or identify cities where homes are in high demand and sell quickly.

11. Property Map (Scatter Mapbox)

Why: Enables spatial exploration of listings with direct reference to price, size, and type. Combines geographic insights with model data, helping visualize market distribution.

Bias and Limitations:

Despite achieving strong model performance, several biases and limitations exist in my project. Firstly, the data set is heavily concentrated on urban and suburban properties, potentially underrepresenting rural markets.

Secondly, in real-world, property prices are influenced by several other external and important factors like crime rates, school quality, and neighborhood amenities, weather, offices near homes, etc which are not captured in the current features, introducing omitted variable bias.

Thirdly, geographic sampling bias may arise because certain states (mainly in the pacific region) dominate the dataset. Lastly, while Random Forest and XGBoost are powerful, they can be overfit to subtle noise if the feature set is not comprehensive. My future enhancements would be to incorporate richer location-based features and external socio-economic indicators to improve fairness and generalizability.

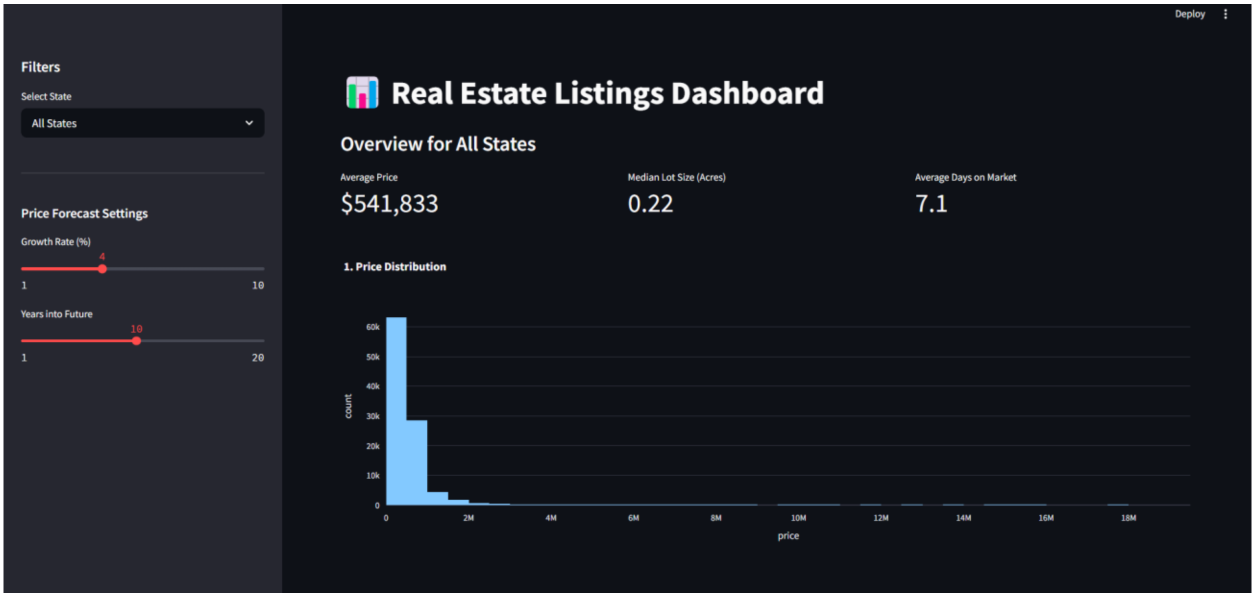
How to start Dashboard?

```
Cd Scripts
```

```
streamlit run Milestone3.py
```

INTERACTIVE DASHBOARD PLOTS:

- **Filter selected: ALL STATES**



Filters

Select State: All States

Price Forecast Settings

Growth Rate (%): 4

Years into Future: 10

Real Estate Listings Dashboard

Overview for All States

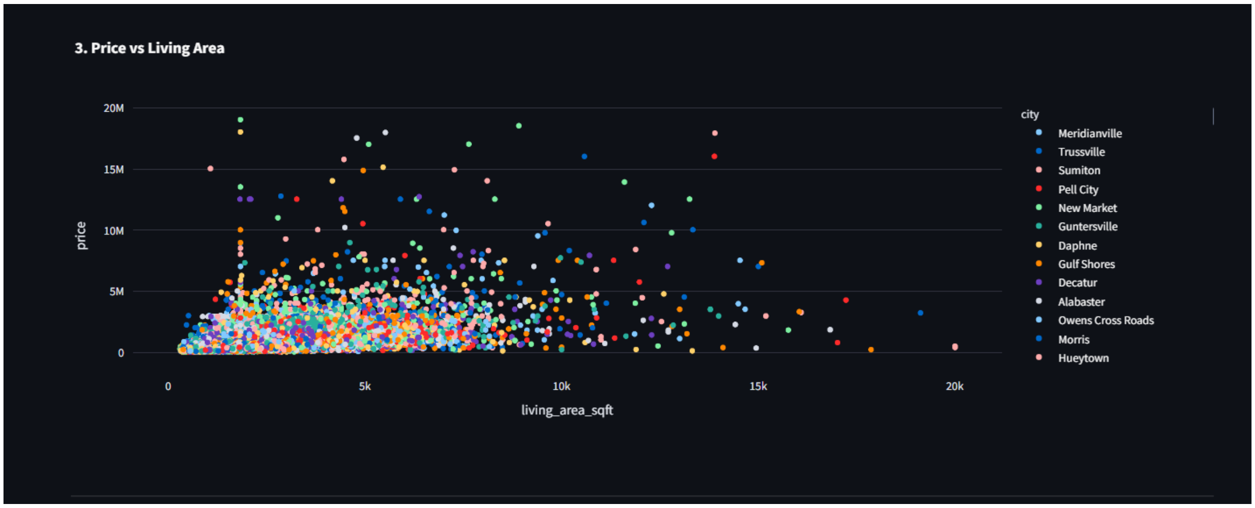
Average Price: $541,833

Median Lot Size (Acres): 0.22

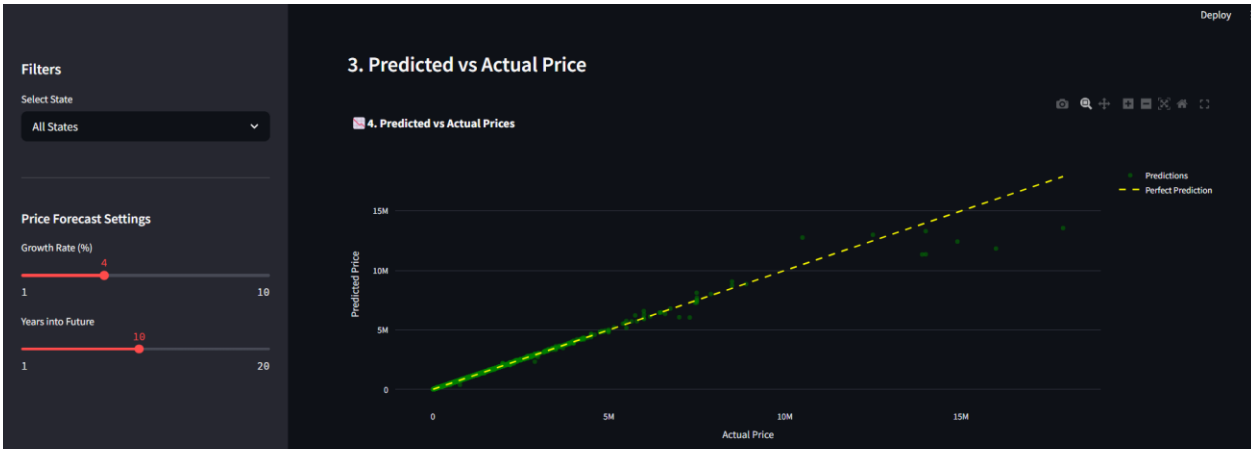
Average Days on Market: 7.1

1. Price Distribution

This histogram shows us the overall average price of all the real-estate data of all the states of America that we have, the median lot size and the average days on market of the properties. It shows the distribution of property prices, helping identify skewness, outliers, and whether the data transformation (e.g., log) is appropriate for stabilization during modeling.



Here you see the cities of all the states of the United States of America. You can also select few of them to visualize the properly or select a specific state and look at the cities of that state only. This gives the estimate of which cities are having expensive real estate in the country or state. It helps visualizes the correlation between living area (square footage) and property price, validating expected positive relationships and detecting outlier homes (extremely high price per sqft).



Filters

Select State: All States

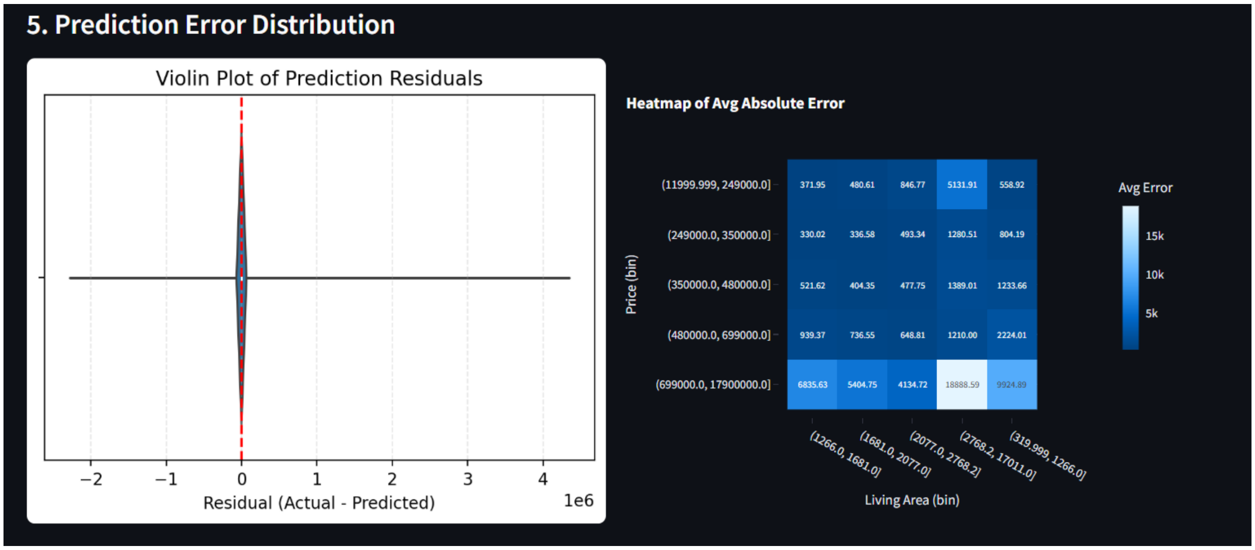
Price Forecast Settings

Growth Rate (%): 4 (1–10)

Years into Future: 10 (1–20)

3. Predicted vs Actual Price

Evaluates model prediction accuracy by comparing predicted prices to actual prices. A perfect model would have all points lying on the diagonal reference line.

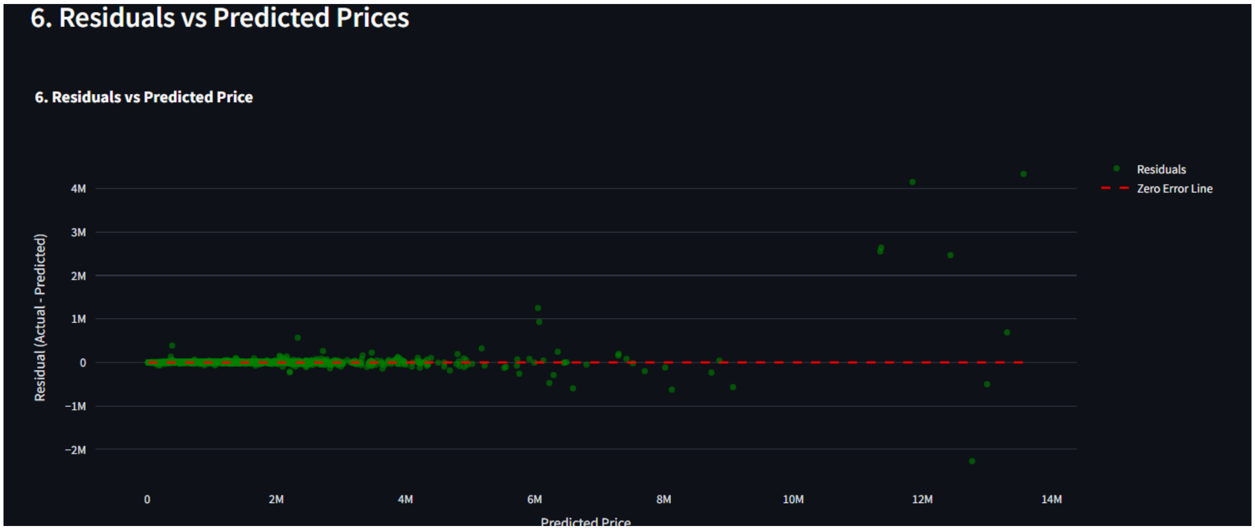


5. Prediction Error Distribution

Price (bin)	(1266.0, 1681.0]	(1681.0, 2077.0]	(2077.0, 2768.2]	(2768.2, 17011.0]	(319.999, 1266.0]
(11999.999, 249000.0]	371.95	480.61	846.77	5131.91	558.92
(249000.0, 350000.0]	330.02	336.58	493.34	1280.51	804.19
(350000.0, 480000.0]	521.62	404.35	477.75	1389.01	1233.66
(480000.0, 699000.0]	939.37	736.55	648.81	1210.00	2224.01
(699000.0, 17900000.0]	6835.63	5404.75	4134.72	18888.59	9924.89

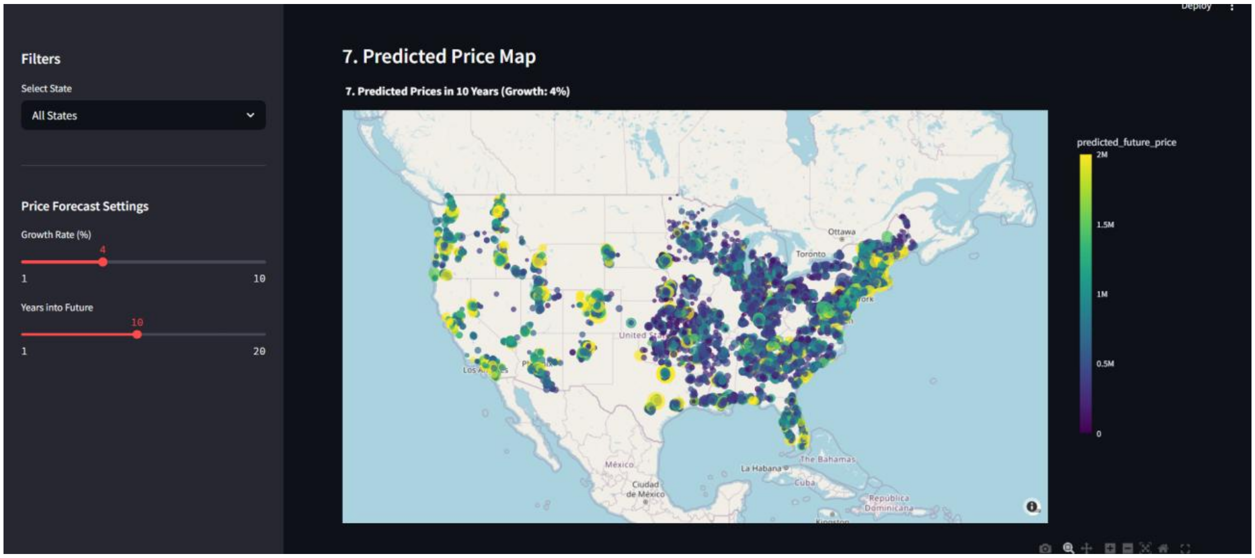
Violin Plot: Visualizes the distribution and spread of residuals (actual - predicted prices) to detect model bias or systematic errors.

Heatmap: Shows how prediction errors vary with respect to different bins of property size and price, helping identify model weakness in specific regions.



6. Residuals vs Predicted Prices

Plots residuals against predicted prices to check for patterns like heteroscedasticity, indicating if model error variances increase with price magnitude.



Filters

Select State

All States

Price Forecast Settings

Growth Rate (%)

4

1 10

Years into Future

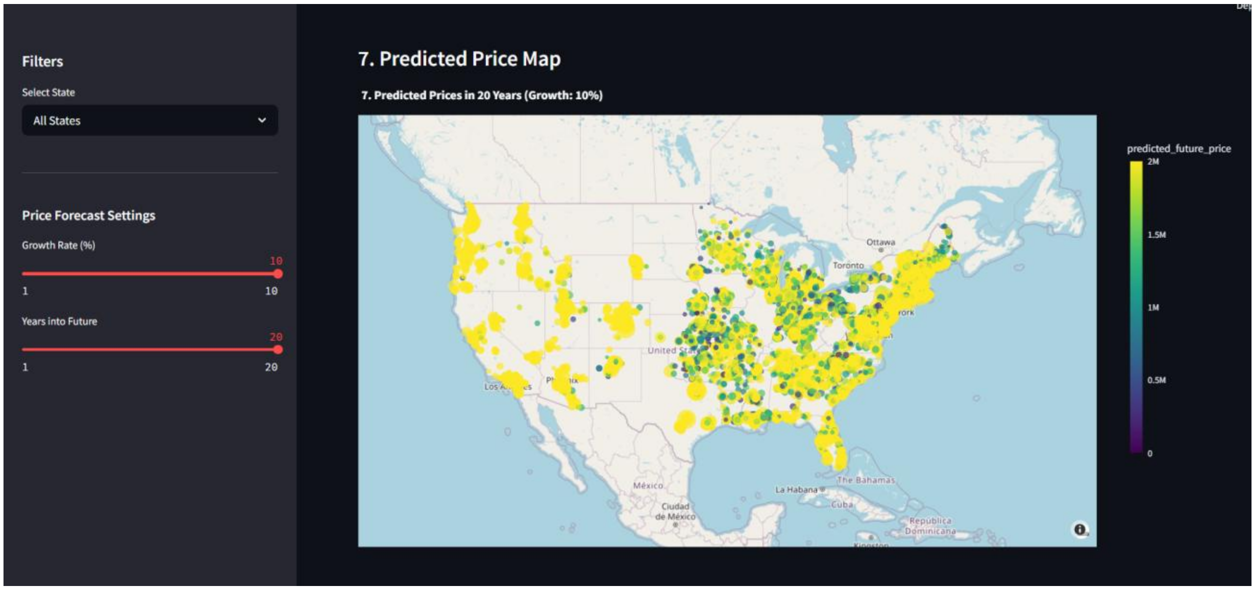
10

1 20

7. Predicted Price Map

7. Predicted Prices in 10 Years (Growth: 4%)

For all states -> Geospatial visualization showing future predicted property prices across the United States based on user-defined growth rate and years into the future, offering a dynamic market forecast view. This is when the growth rate is 4% and the year into future is 10.



Filters

Select State

All States

Price Forecast Settings

Growth Rate (%)

10

1 10

Years into Future

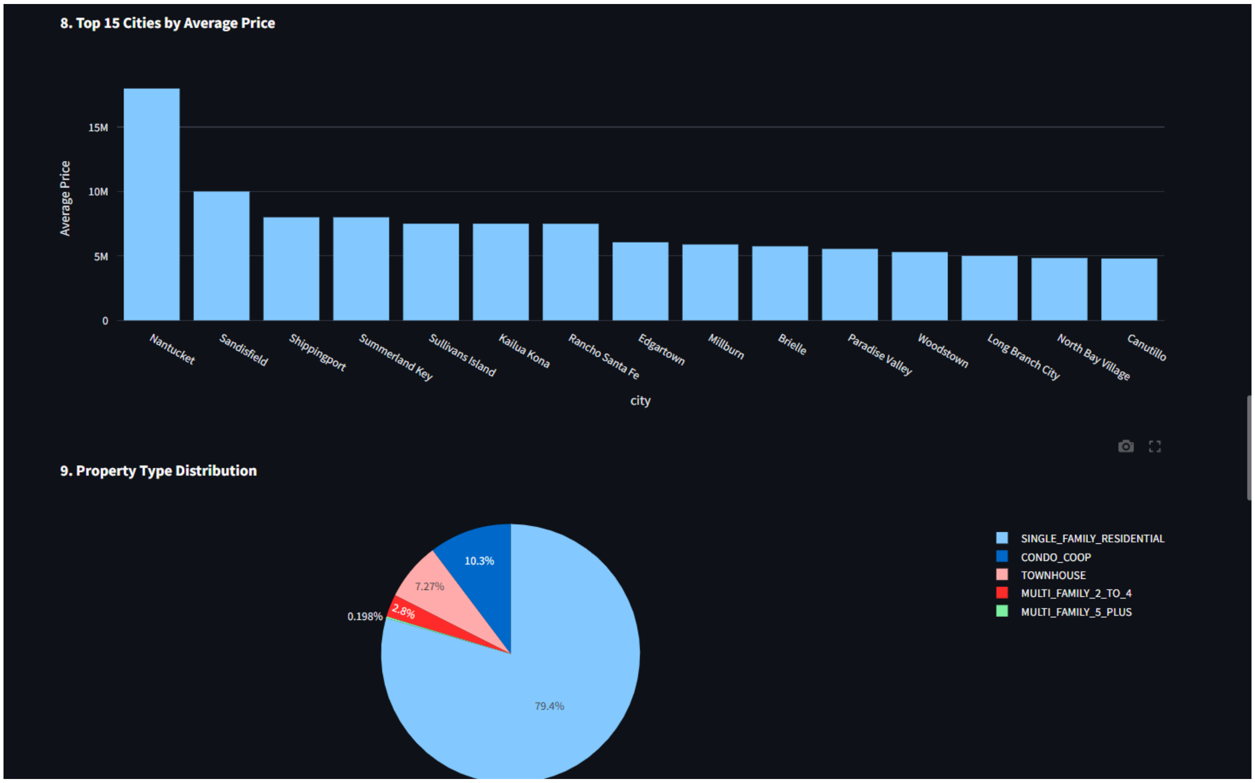
20

1 20

7. Predicted Price Map

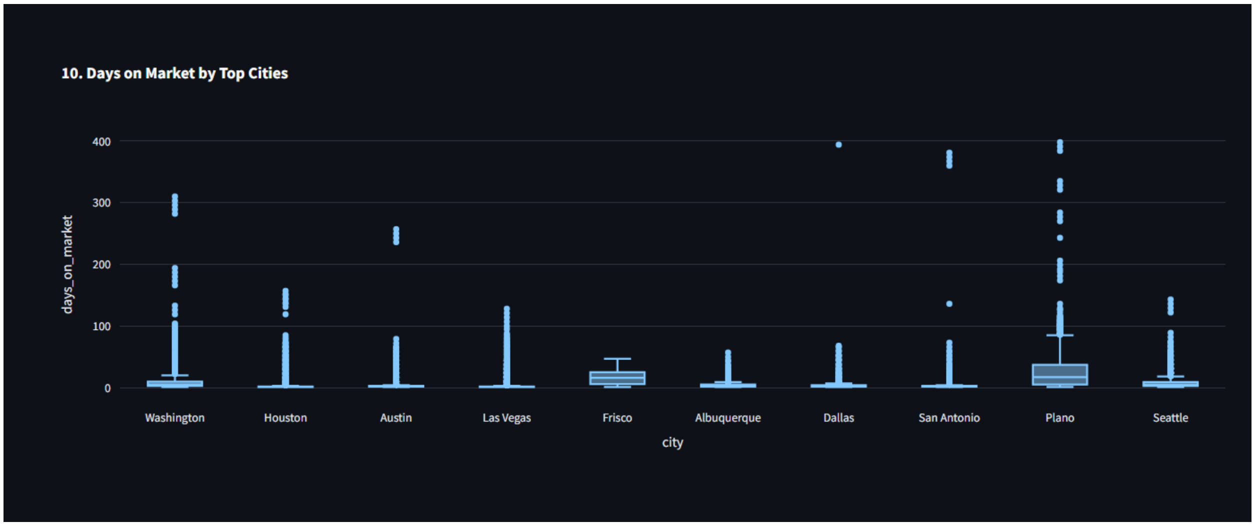
7. Predicted Prices in 20 Years (Growth: 10%)

The same changes when we change the slider for growth rate and years into future. Now you see, for growth % = 10 and years into future = 20, the range of prices are mostly above 1.5 million where as it was in range of 0.5 million for growth rate = 4 and years into future = 10.

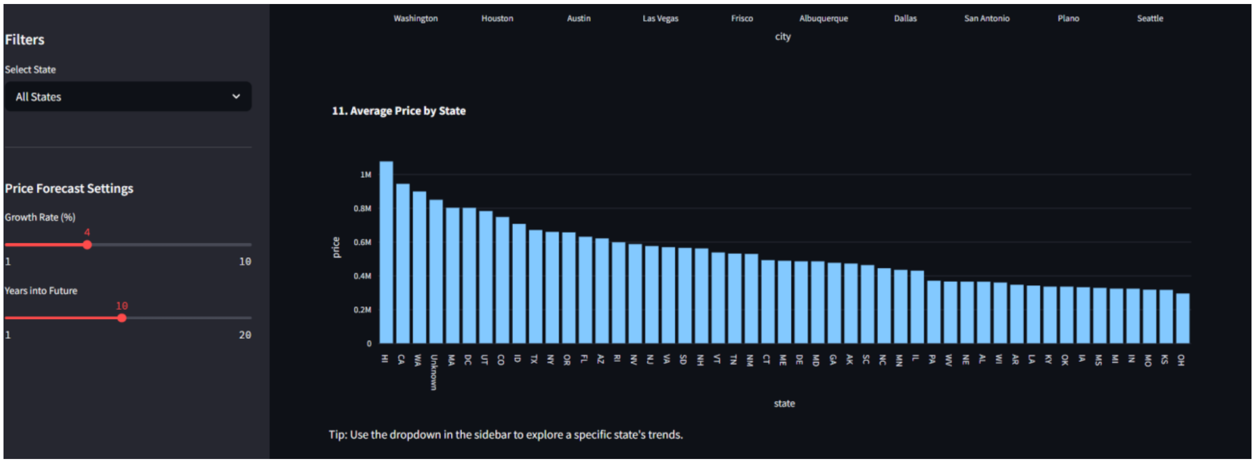


Ranks cities based on average listing price, highlighting regional price disparities and investment opportunities across urban areas.

Displays the breakdown of different property types (single-family, condos, etc.) in the dataset, aiding in understanding inventory composition across markets.



Illustrates variability in how long properties stay listed across major cities, revealing areas with faster-selling or slower-moving real estate markets.



Filters

Select State

All States

Price Forecast Settings

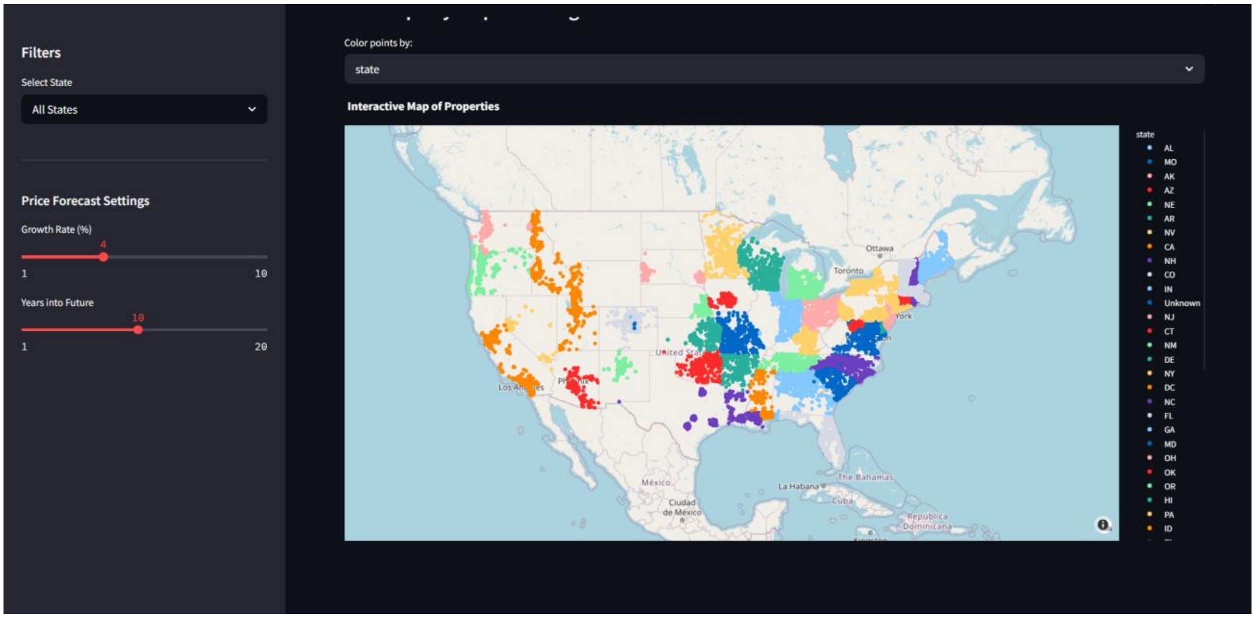
Growth Rate (%): 4

Years into Future: 10

11. Average Price by State

Tip: Use the dropdown in the sidebar to explore a specific state's trends.

Compares mean property prices across states, helping users detect state-level trends and relative affordability or premium market regions.



Filters

Select State

All States

Price Forecast Settings

Growth Rate (%): 4

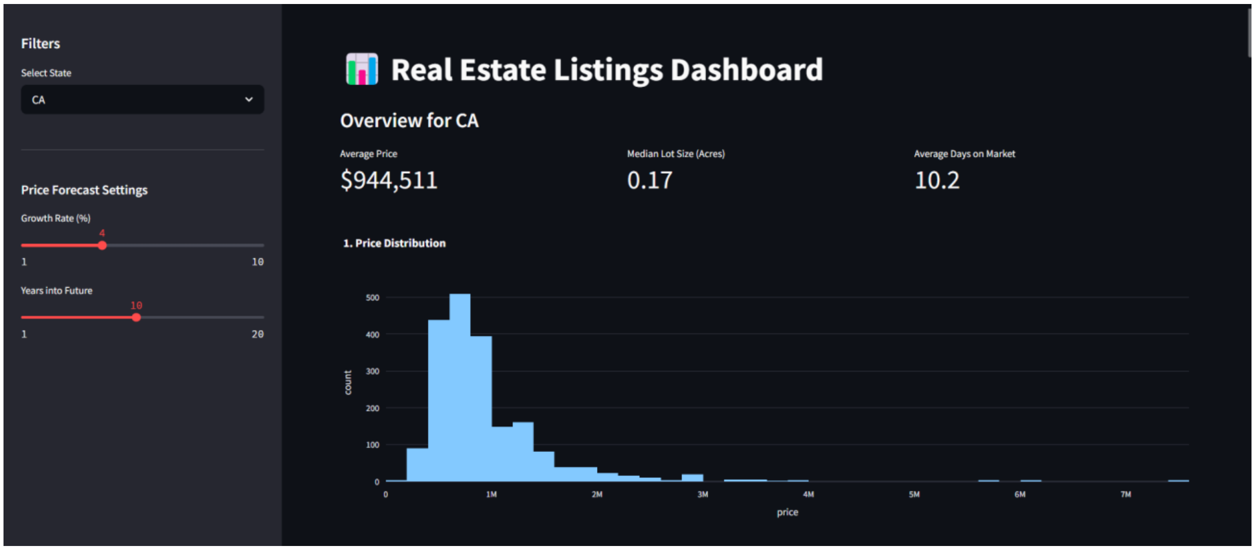
Years into Future: 10

Color points by: state

Interactive Map of Properties

Plots all active property listings on an interactive map, colored by selected attributes (e.g., price, days on market), to reveal geographic price clusters and inventory density.

Filter State : CA (California)



Filters

Select State: CA

Price Forecast Settings

Growth Rate (%): 4 (1 – 10)

Years into Future: 10 (1 – 20)

Real Estate Listings Dashboard

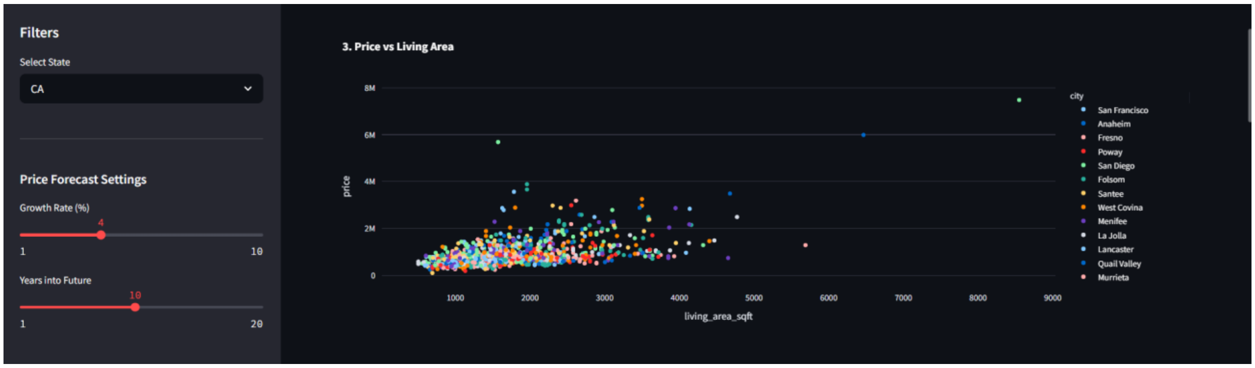
Overview for CA

Average Price: $944,511

Median Lot Size (Acres): 0.17

Average Days on Market: 10.2

1. Price Distribution



Filters

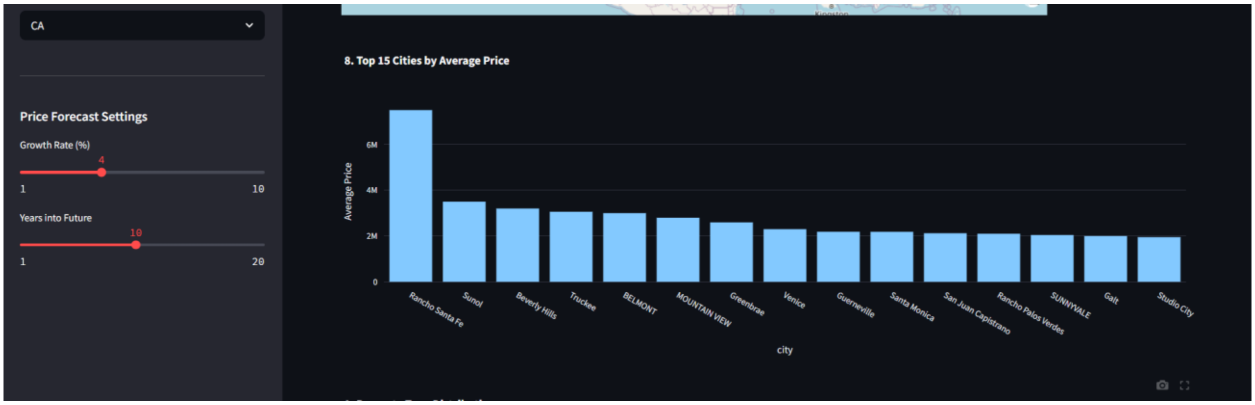
Select State: CA

Price Forecast Settings

Growth Rate (%): 4 (1 – 10)

Years into Future: 10 (1 – 20)

3. Price vs Living Area



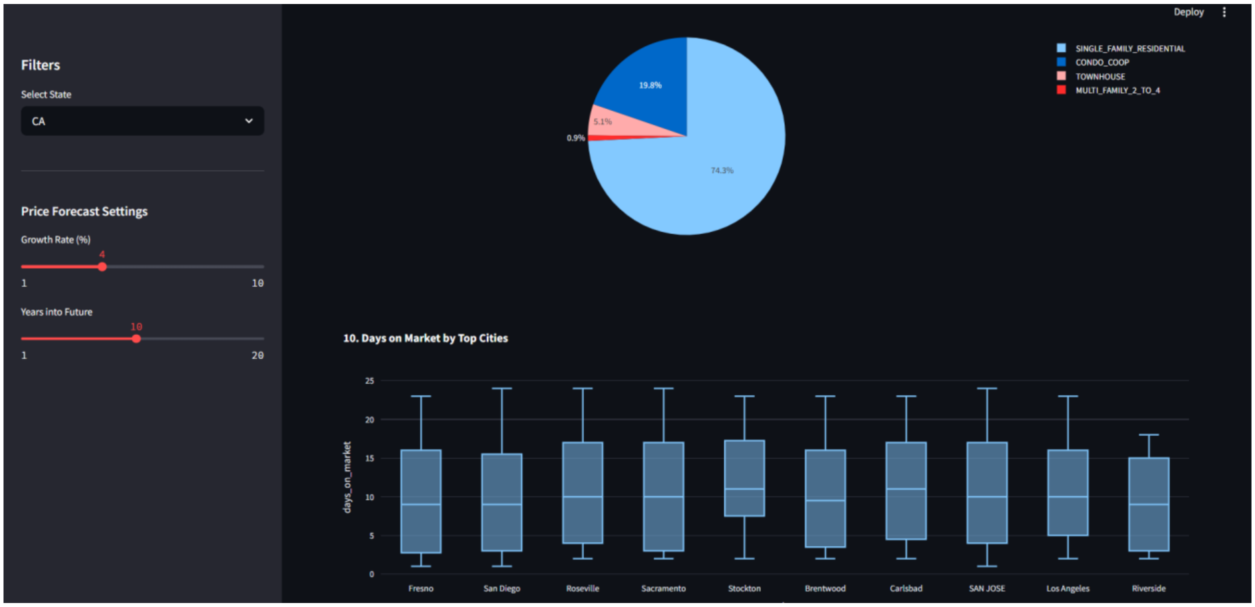
CA

Price Forecast Settings

Growth Rate (%): 4 (1 – 10)

Years into Future: 10 (1 – 20)

8. Top 15 Cities by Average Price



Filters

Select State

CA

Price Forecast Settings

Growth Rate (%)

4

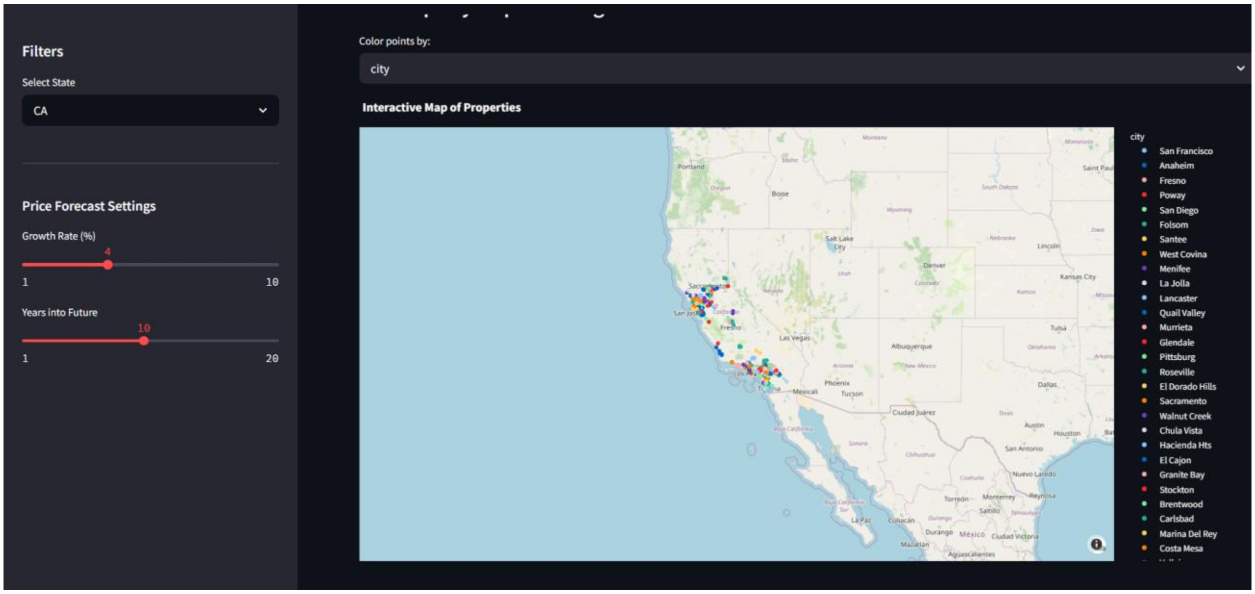
1 10

Years into Future

10

1 20

10. Days on Market by Top Cities



Filters

Select State

CA

Price Forecast Settings

Growth Rate (%)

4

1 10

Years into Future

10

1 20

Color points by:

city

Interactive Map of Properties

Tech Stack:

Category	Technologies / Libraries	Purpose
Programming Language	Python 3.x	Main language for development
Data Manipulation	pandas, numpy	Data cleaning, transformation, feature engineering
Machine Learning	scikit-learn (Linear Regression, Random Forest Regressor, train_test_split, StandardScaler, metrics like MAE, RMSE, R^2, ROC AUC)	Model building, training, evaluation
Visualization	matplotlib, seaborn, plotly.express, plotly.graph_objects	Static and interactive data visualizations (histograms, scatter plots, bar charts, violin plots, residual plots, maps)
Dashboard Development	Streamlit	Building the interactive real estate dashboard application
Report Generation	reportlab	Creating and exporting evaluation reports as PDF files
Utility Libraries	os, json	File system operations and saving/loading model evaluation results

LLM Usage Declaration:

I used ChatGPT to clarify report structure, check grammar and refine technical writing for the report. All outputs were critically reviewed and validated by me.

Video Submissions:

Milestone 1 and 2 Video -> https://www.youtube.com/watch?v=Aq4uK_sWo9E

Milestone 3 Video (Tool Video) -> https://www.youtube.com/watch?v=0PViKMird1Q

Justification for Scope Reduction:

After discussing with Dr. Grant regarding the scope of my project, which included analyzing the relationships between gold, real estate, and the stock market — was too broad for the course timeline and deliverable structure and I found it difficult to find proper relation between these 3 different data. Therefore, I formally requested and received approval to reduce the scope to focus exclusively on the real estate market, allowing for deeper analysis, more refined modeling, and the development of a meaningful and interactive dashboard tool. Here is the link of my work on my pervious project which was if much larger scope

Link -> https://github.com/SaiPande/IDS-Previous-Submission