Predictive Modeling and Visualization of U.S. Rental Trends Using Zillow Rent Index

Problem Statement

The U.S. rental housing market has experienced substantial volatility over the past decade, driven by economic fluctuations, demographic shifts, and policy changes. Rising rents and uneven regional growth have intensified affordability challenges, particularly in rapidly expanding urban and suburban regions. These dynamics affect multiple stakeholders, including renters seeking stable housing, policymakers making zoning and development decisions, and real estate investors assessing risk and opportunity.

Zillow's Observed Rent Index (ZORI) provides comprehensive historical rent data across U.S. cities and ZIP codes. However, because the dataset is retrospective and not inherently structured for forecasting, stakeholders lack accessible, data-driven tools that can generate forward-looking insights at granular geographic levels.

This project addresses that gap by leveraging historical ZORI data to build predictive models capable of forecasting median rent prices over a 12-month horizon at the city level.

Data Wrangling

The data wrangling phase focused on converting the Zillow Observed Rent Index (ZORI) dataset into a clean, structured format suitable for exploratory analysis, time-series modeling, and visualization.

The raw dataset contains monthly median rent values across multiple geographic levels, including cities, ZIP codes, counties, and metropolitan areas. The data is structured in a wide format, with each column corresponding to a specific month and year. Initial inspection included reviewing data types, verifying consistency across columns, and identifying missing or anomalous values.

To align the dataset with project objectives, the analysis centered on data at the city and ZIP code levels. Non-essential fields were removed, while key identifiers—such as city, state, and metropolitan area—were retained to support regional comparisons.

Time-series modeling requires chronological structure, so the dataset was reshaped from wide to long format. Monthly rent columns were unpivoted into a single date column paired with corresponding rent values. A unified date field was then created to ensure proper chronological indexing.

Missing values were addressed through a combination of removing records with insufficient data when appropriate.

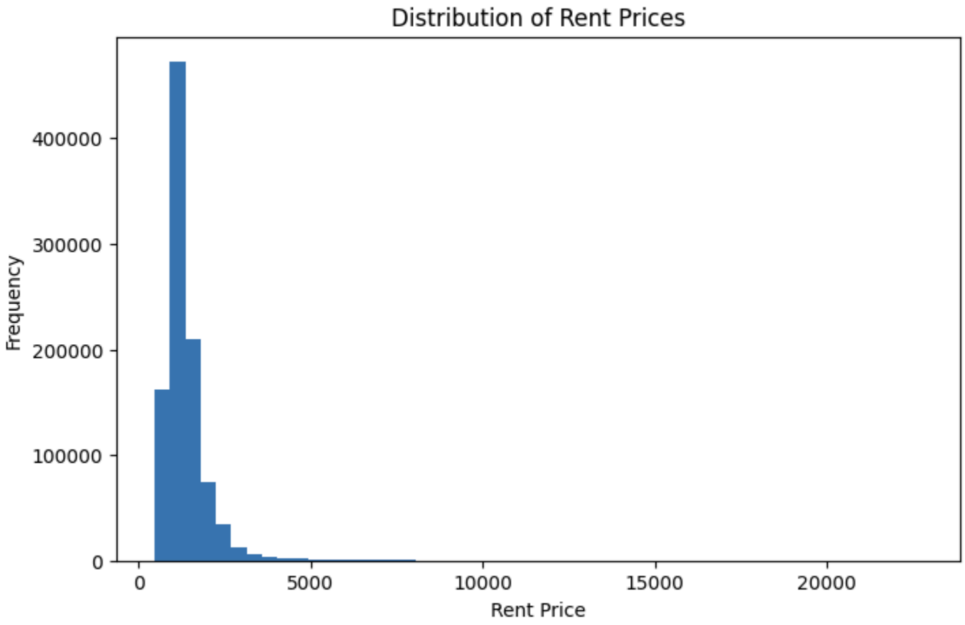
This method preserved the continuity of the time-series data while minimizing distortions.

Outliers in rent prices were identified through statistical thresholds. Extreme values were capped to reduce the influence of anomalous spikes that could negatively impact model training and forecast stability. This step helped preserve overall trend behavior while improving robustness.

EDA

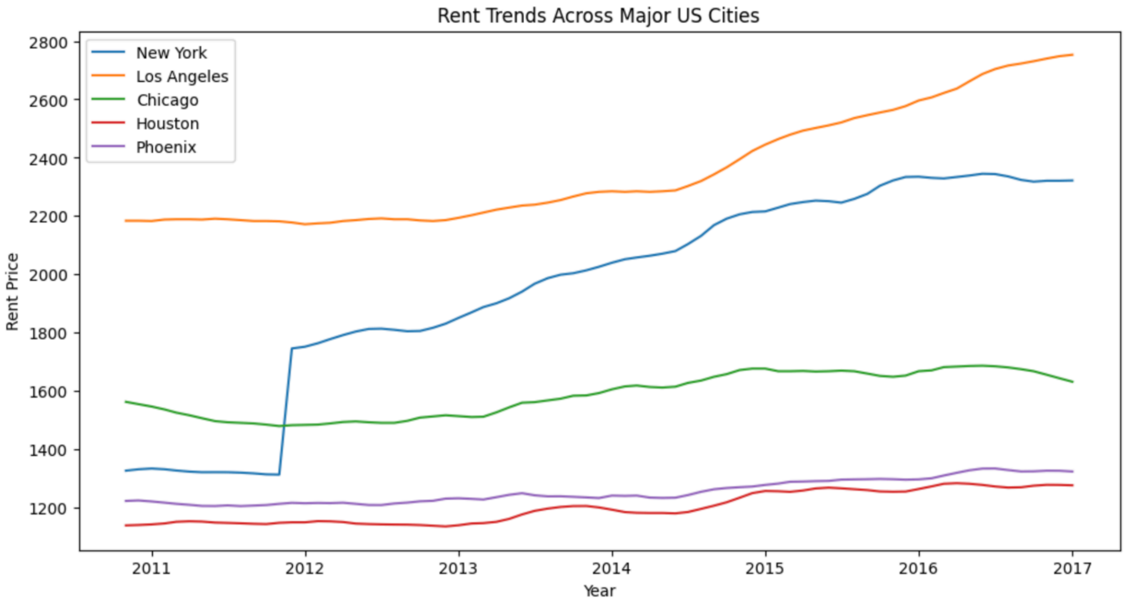
EDA was conducted to understand the distribution of rental prices, assess historical trends, and uncover relationships between key variables.

Median rent prices exhibited a right-skewed distribution, with most regions concentrated in the lower-to-mid price ranges and a long tail of high-rent areas.



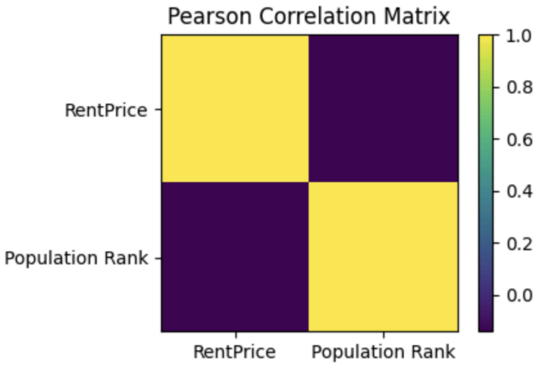
This pattern aligns with known market behavior and confirms that outlier adjustments preserved meaningful variation.

Average monthly rent prices were analyzed for major U.S. cities.



All cities demonstrated upward trajectories, though the magnitude and rate of increase varied significantly. Larger metropolitan areas consistently showed higher baseline rents, reinforcing the importance of geographic context in forecasting models.

A positive though moderate relationship was observed between population rank and rent levels. While larger cities tend to have higher rents, population alone does not adequately explain rent variability. This confirmed the need for models integrating both temporal and geographic features.



Overall, EDA validated the presence of significant temporal and spatial patterns and informed subsequent modeling strategies.

Preprocessing and Data Training

Following EDA, the dataset was further preprocessed to prepare for model development.

Temporal features such as year and month were retained to capture long-term trends and seasonal cycles. Geographic attributes such as city and state were encoded numerically to support machine learning algorithms. Numerical features were scaled when necessary to prevent disproportionate influence from features with larger magnitudes.

The dataset was divided chronologically into training and testing subsets to simulate real-world forecasting scenarios. Only past data was used to predict future values, ensuring proper evaluation.

Modeling

The goal of the modeling phase was to develop a robust forecasting approach for rental prices.

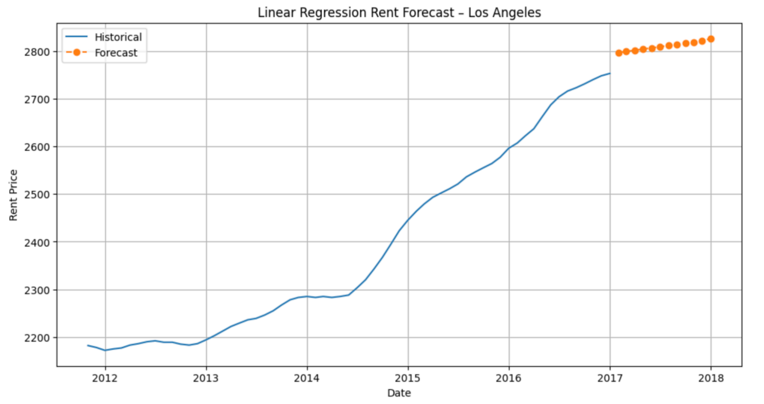
Since the original dataset consisted primarily of dates and rental values, additional temporal features were engineered. The date column was decomposed into year and month, and an ordinal time index was created to capture trend components.

A Linear Regression model was trained using scikit-learn. Performance on the test set yielded:

- **MAE:** 254.73
- **RMSE:** 495.08
- **R^2:** 0.4328

This indicates that the model explained approximately 43% of the variance in rental prices.

Forecasts were then generated for a sequence of future dates using the same feature engineering pipeline.



Random Forest and Gradient Boosting models were also evaluated. Their performance metrics were **Random Forest:** MAE 346.80, RMSE 594.90, R² 0.1811 - **Gradient Boosting:** MAE 300.15, RMSE 541.82, R² 0.3207

Although these models are more complex, they underperformed compared with linear regression and required substantially greater computational resources. Linear Regression offered the best balance of accuracy, interpretability, and efficiency, making it the preferred model for this task.

Interactive Dashboard

An interactive dashboard was developed using Streamlit to provide stakeholders with an accessible and dynamic interface for exploring both historical rent data and model-generated forecasts. The dashboard integrates the cleaned Zillow rent dataset and the trained Linear Regression model to deliver city-level insights. Users can select a city from the sidebar, view historical rent trends, and examine a 12-month forecast displayed through intuitive Plotly visualizations. The dashboard also computes key metrics such as the latest actual rent value and the forecast for the upcoming month, enabling quick comparisons and decision support. By combining interactivity with real-time computation, the dashboard transforms the forecasting model into a practical, user-friendly tool for renters, policymakers, analysts, and real estate stakeholders.

Conclusion

This project demonstrates that historical ZORI data can be effectively transformed and modeled to generate short-term rental price forecasts. Through comprehensive data wrangling, exploratory analysis, and model evaluation, Linear Regression emerged as the most reliable and interpretable approach for predicting city-level rent trends. The integration of these forecasts into an interactive Streamlit dashboard further enhances the project's practical value, allowing users to visualize trends, compare historical and predicted values, and make informed decisions. Overall, the project provides a scalable and transparent framework for understanding housing market dynamics and offers a foundation for future enhancements, including incorporation of additional socioeconomic variables or advanced modeling techniques.