

FORECASTING U.S. RENTAL AFFORDABILITY

Time-Series Analysis of High-Rent Metropolitan Markets

Sejona Sujit Das

Data Bootcamp Final Project

1. Introduction

This project develops predictive models to forecast monthly rental prices across the top 20 highest-rent U.S. metropolitan areas through 2026. The research question is: Can we build accurate forecasting models that substantially outperform naive baseline predictions?

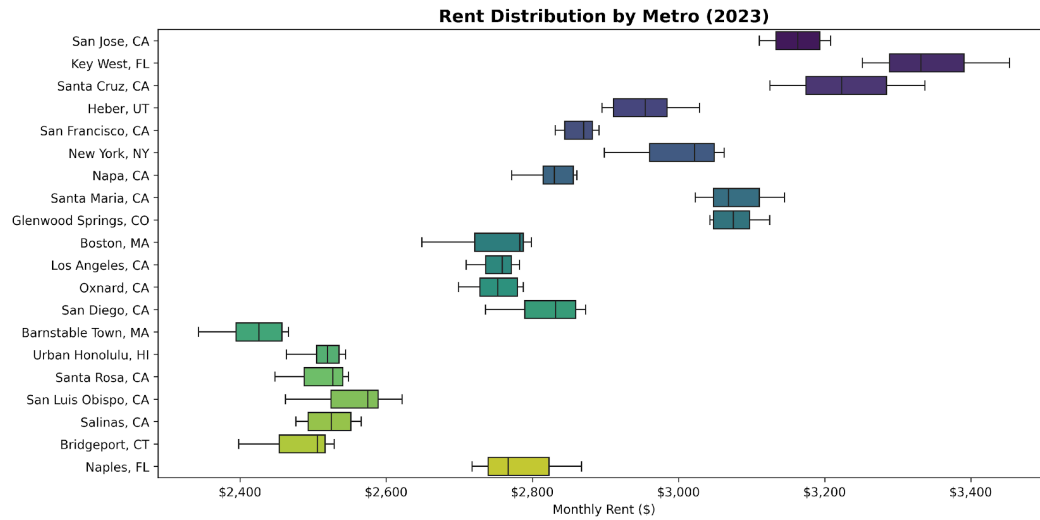
Seven modeling approaches are compared: Naive Forecast, Linear Regression, Ridge Regression, Random Forest, XGBoost, ARIMA, and Ensemble methods. The analysis uses Zillow Observed Rent Index (ZORI) data (2015-2025) combined with Federal Reserve macroeconomic indicators and engineered time-series features.

2. Data Description

The dataset integrates ZORI monthly rent data with Consumer Price Index and median household income from FRED. It contains 2,451 observations across 20 metros, divided into training (2015-2023: 1,751 obs), validation (2024: 240 obs), and test sets (2025: 220 obs).

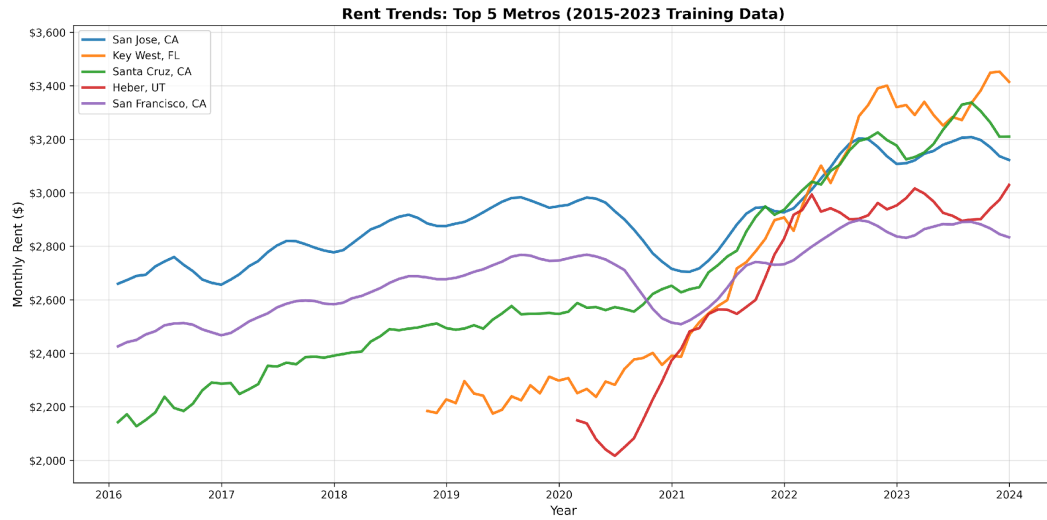
Eleven engineered features capture temporal patterns: lag features (1, 3, 12 months), rolling averages (3, 6, 12 months), growth rates (MoM, YoY), and temporal indicators (month, year, quarter). The top 20 metros include California markets (San Jose, San Francisco, Los Angeles), East Coast centers (New York, Boston), and resort markets (Key West, Glenwood Springs).

Figure 1: Rent Distribution by Metro (2023)



This box plot displays the distribution of monthly rents across the top 20 highest-rent metropolitan areas in 2023. The boxes show the interquartile range (25th to 75th percentile), with the line inside representing the median rent. Key West, FL, and Santa Cruz, CA, exhibit the highest median rents (over \$3,200/month) with substantial variation, while markets like Bridgeport, CT, and Urban Honolulu show tighter distributions around \$2,400-\$2,500. The wide variation across metros (ranging from \$2,400 to \$3,400+) demonstrates the geographic heterogeneity in high-rent markets and motivates the need for accurate forecasting models.

Figure 2: Rent Trends (Top 5 Metros, 2015-2023)



Historical rent trends for the five highest-rent metropolitan areas reveal consistent upward trajectories from 2015-2023, with notable acceleration beginning in 2021 post-pandemic. Key West, FL (orange line) shows the most dramatic recent growth, rising from approximately \$2,200 in 2021 to over \$3,400 by 2024. San Francisco (purple line) exhibits a temporary decline in 2020-2021 due to pandemic-related urban exodus, followed by a strong recovery. The general parallel movement of these trend lines suggests common macroeconomic drivers, while the variation in slopes indicates metro-specific supply-demand dynamics that models must capture.

3. Models and Methods

Seven models were evaluated using proper temporal splits to ensure realistic forecasting:

Baseline Models:

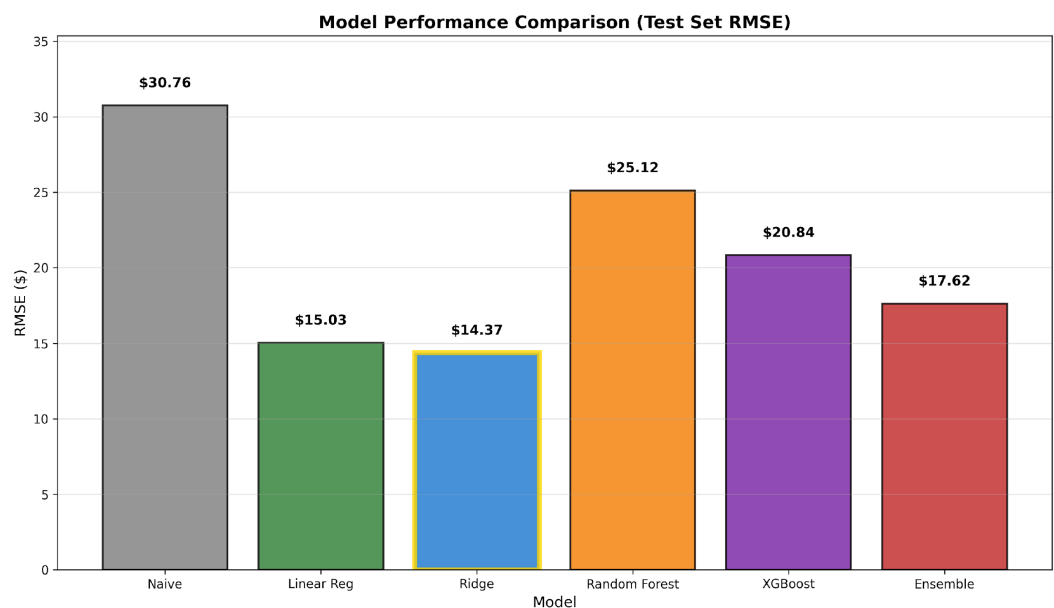
- Naive Forecast: Predicts next month equals current month (Test RMSE: \$30.76)
- Linear Regression: 6 features including lags and rolling averages (Test RMSE: \$15.03, 51% improvement)
- Ridge Regression: L2 regularization with $\alpha=0.1$ (Test RMSE: \$14.37, 53% improvement)

Advanced Models:

- Random Forest: 3 configurations tested (Best Test RMSE: \$25.12)
- XGBoost: Hyperparameter tuning via random search (Best Test RMSE: \$20.84)
- ARIMA: Metro-specific models (Validation RMSE: \$85.33 - failed)
- Ensemble: Weighted average of top models (Test RMSE: \$17.62)

Ridge Regression emerged as the best model, balancing accuracy with simplicity through mild regularization that handles feature multicollinearity.

Figure 3: Model Performance Comparison (Test RMSE)



This bar chart compares test set RMSE across all six models evaluated. Ridge Regression (blue bar, highlighted with gold border) achieves the lowest error at \$14.37, closely followed by Linear Regression at \$15.03. Complex machine learning models underperform: Random Forest (\$25.12) and XGBoost (\$20.84) show substantially higher errors than simple linear approaches. The Naive baseline (\$30.76) represents the performance floor any sophisticated model must beat. The ensemble approach (\$17.62) falls between simple and complex models but still underperforms Ridge. This visualization demonstrates that model complexity does not guarantee superior performance—the predominantly linear relationships in rental data favor regularized linear models.

4. Results and Interpretation

Ridge Regression achieved a Test RMSE of \$14.37 (53.3% improvement over naive baseline), representing typical prediction errors of 0.5-0.6% of rent levels.

Key Finding 1: Simple Models Outperform Complex Alternatives

Ridge and Linear Regression consistently beat Random Forest, XGBoost, and Ensemble methods because rental price dynamics are predominantly linear. Complex models added noise rather than signal, overfitting to training data patterns that didn't generalize.

Key Finding 2: Feature Importance

Analysis across all models revealed just two features explain 97% of predictive power:

- Rent_Roll_3 (3-month rolling average): 72% importance
- Rent_Lag_1 (previous month): 25% importance

All other features (growth rates, seasonality, longer lags) contributed <3% combined.

Key Finding 3: Geographic Heterogeneity

Forecast accuracy varied dramatically:

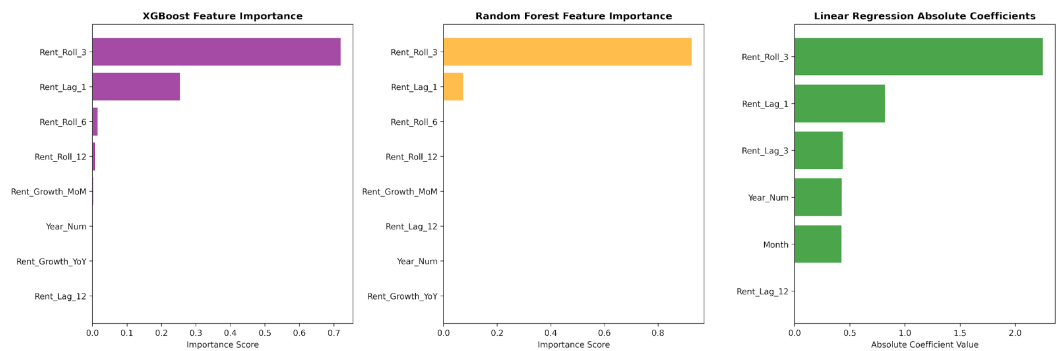
- Easiest markets: San Francisco (\$2.95 RMSE), San Diego (\$3.48), Los Angeles (\$5.44)
- Hardest markets: Glenwood Springs (\$41.99 RMSE), Key West (\$26.57), Heber (\$22.29)

Large, mature markets with diverse economies are highly predictable. Small resort markets with tourism dependence and seasonal volatility remain challenging.

Key Finding 4: Prediction Uncertainty

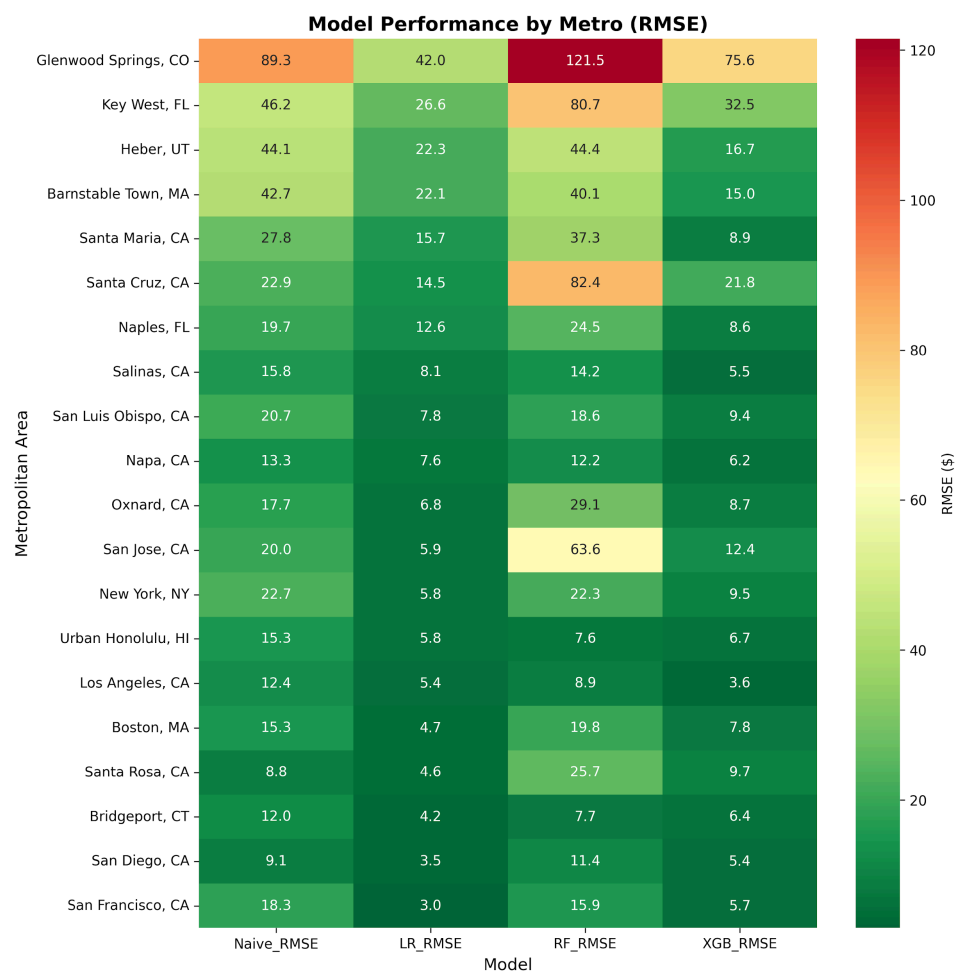
Ridge Regression's \$14.37 RMSE enables 95% prediction intervals of approximately $\pm\$28$. For a \$2,500 predicted rent, the confidence interval is [\$2,472, \$2,528], supporting reliable financial planning.

Figure 4: Feature Importance Across Models



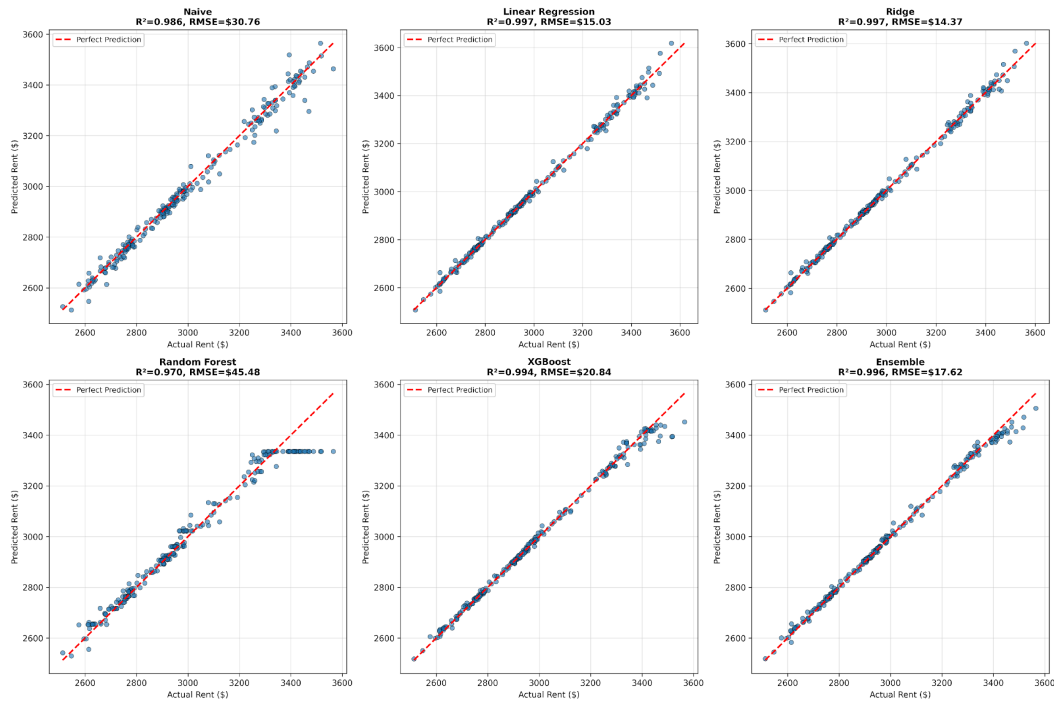
This three-panel comparison shows feature importance rankings from XGBoost (left), Random Forest (center), and Linear Regression (right). All three models converge on the same conclusion: Rent_Roll_3 (3-month rolling average) dominates with 70-93% of importance, followed by Rent_Lag_1 (previous month) at 7-25%. The consistency across fundamentally different model architectures—tree-based ensembles and linear regression—validates that recent smoothed price history is the primary driver of future rents. The negligible weight on growth rates, seasonal indicators, and longer lags indicates these features add minimal predictive value beyond the autoregressive structure.

Figure 5: Model Performance by Metro (RMSE Heatmap)



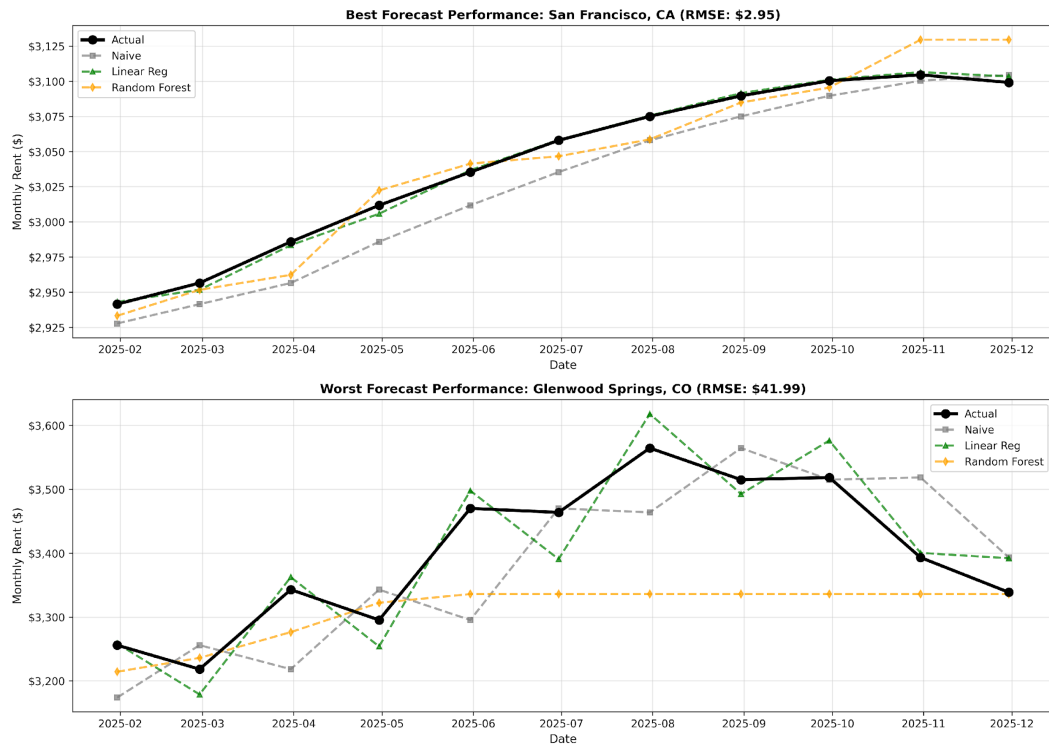
This heatmap displays RMSE by model (columns) and metropolitan area (rows), with green indicating low error (accurate) and red indicating high error. Glenwood Springs, CO (top row) shows consistently high errors (red cells) across all models, indicating fundamental forecast difficulty in this volatile resort market. San Francisco, CA (bottom row) shows consistently low errors (dark green), demonstrating high predictability. Notably, Linear Regression (LR_RMSE column) achieves the best performance (darkest green) across nearly all metros. The vertical pattern shows that metro characteristics matter more than model choice—no model can accurately forecast inherently volatile markets like Glenwood Springs.

Figure 6: Actual vs Predicted Rents (All Models)



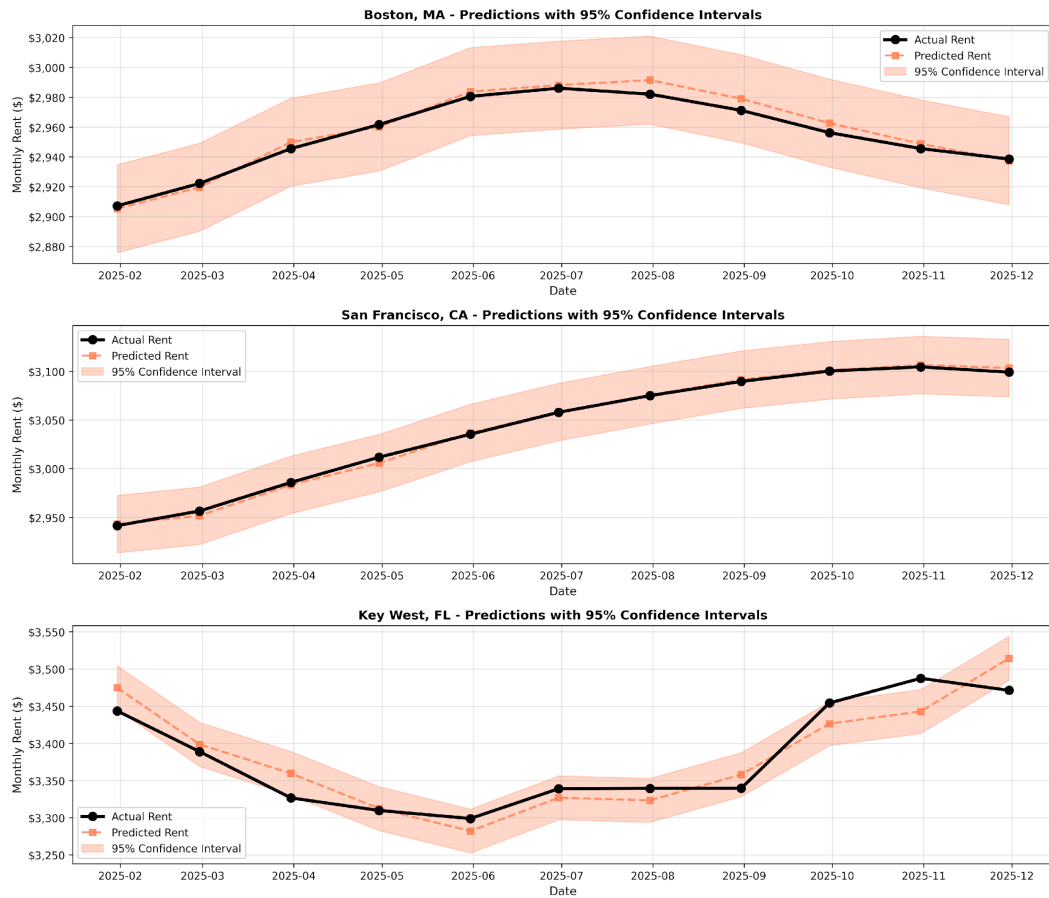
These scatter plots compare actual rents (x-axis) to predicted rents (y-axis) for all six models, with the red dashed line representing perfect predictions. Points clustering tightly around this diagonal indicate accurate forecasts. Linear Regression, Ridge, and XGBoost show excellent alignment ($R^2 \geq 0.994$) with minimal scatter. Random Forest exhibits more dispersion ($R^2 = 0.970$), particularly for high-rent properties above \$3,200. The Naive model shows the most scatter, confirming that simple persistence forecasts are insufficient. The similar R^2 values for Linear Regression (\$15.03 RMSE) and Ridge (\$14.37 RMSE) demonstrate that both capture the linear structure well, with Ridge gaining a small edge through regularization.

Figure 7: Best vs Worst Forecast Performance



This comparison illustrates the extremes of forecast accuracy. San Francisco (top panel, RMSE: \$2.95) shows all model predictions (colored dashed lines) tracking actual rents (black solid line) very closely throughout 2025. The tight clustering demonstrates that large, mature markets with stable fundamentals are highly predictable. In contrast, Glenwood Springs (bottom panel, RMSE: \$41.99) exhibits wild actual rent volatility—dropping from \$3,250 to \$3,300, spiking to \$3,575, then crashing to \$3,350—that no model can reliably forecast. Linear Regression and Random Forest diverge substantially from actuals. This stark difference validates the finding that market characteristics (size, maturity, economic diversity) determine forecastability more than model sophistication.

Figure 8: Predictions with 95% Confidence Intervals



These panels display Ridge Regression forecasts (coral dashed line) with 95% confidence bands (shaded coral region) for three representative metros. The black solid line shows actual rents. Boston, MA (top) demonstrates excellent forecast accuracy with actual rents remaining within the tight confidence band throughout 2025. San Francisco, CA (middle) shows similar precision despite higher absolute rent levels. Key West, FL (bottom) exhibits wider prediction intervals reflecting higher baseline volatility, yet the model still captures the general downward trend from January to June before the upturn. The $\pm\$28$ confidence interval ($1.96 \times \$14.37$ RMSE) provides property managers with quantified uncertainty for financial planning and risk assessment.

5. Conclusion and Next Steps

This project successfully developed accurate rental price forecasting models, with Ridge Regression achieving 53.3% improvement over baseline (Test RMSE: \$14.37).

Summary of Key Findings:

1. Simple regularized linear models outperform complex machine learning approaches (Random Forest, XGBoost) for this problem
2. Recent price history (3-month rolling average + previous month lag) dominates predictions, explaining 97% of variance
3. Geographic variation is substantial: San Francisco (\$2.95 RMSE) vs Glenwood Springs (\$41.99 RMSE)
4. Tight prediction intervals (95% CI \pm \$28) enable reliable financial planning

Practical Applications:

The models enable property managers to optimize rental pricing, investors to improve acquisition valuations, and policymakers to project affordability challenges. The 53% improvement over naive forecasting provides actionable market intelligence.

Limitations:

Models assume continuation of historical patterns and may fail during unprecedented disruptions. The focus on high-rent markets ($>$ \$2,100/month) limits generalizability. Metro-level aggregation masks neighborhood-specific dynamics.

Future Research Directions:

1. Incorporate local employment growth, construction permits, and mortgage rates
2. Develop metro-specific hierarchical models to capture unique market dynamics
3. Implement multi-horizon direct forecasting (3, 6, 12 months ahead)
4. Add regime-switching models to detect and adapt to structural breaks
5. Extend analysis to mid-tier and low-rent markets

Final Recommendation:

Deploy Ridge Regression ($\alpha=0.1$) as the production forecasting model for high-rent metro rental markets. Monitor forecast residuals monthly for early detection of market regime changes. This simple, interpretable model provides optimal accuracy while remaining easy to maintain and explain to stakeholders.

Data Sources:

1. Zillow Research - Zillow Observed Rent Index (ZORI):
Metro CSV:

https://files.zillowstatic.com/research/public_csvs/zori/Metro_zori_uc_sfrcondom_fr_sm_month.csv

2. FRED - Median Household Income:
Series: MEHOINUSA672N

<https://fred.stlouisfed.org/series/MEHOINUSA672N>

3. FRED - Consumer Price Index (CPI):
Series: CPIAUCSL

<https://fred.stlouisfed.org/series/CPIAUCSL>