Housing Affordability Analysis and Forecasting Across Pennsylvania Metropolitan Areas

Prepared by:
Colton Dumm and Ryan Quinlan

Abstract
This technical report analyzes housing affordability across Pennsylvania's major metropolitan areas (Philadelphia, Pittsburgh, and Harrisburg) and their constituent counties from 2012 to 2023, with forecasts extending to 2028. Using an integrated dataset combining Redfin housing market statistics with U.S. Census Bureau socioeconomic indicators, we implemented dual modeling approaches: ARIMA (Autoregressive Integrated Moving Average) for time series forecasting and XGBoost for machine learning prediction. Our analysis addresses key economic questions including regional affordability variations, emerging housing cost patterns relative to income levels, potential investment opportunities across different counties, and implications for affordable housing policy initiatives.

The affordability ratio (median sale price/median household income) is the primary metric for evaluating housing market accessibility. ARIMA captures temporal patterns and cyclical market behaviors, while XGBoost identifies complex relationships between socioeconomic factors and housing affordability. Median price per square foot, total population, and educational attainment are the most significant influences on affordability ratios.

The findings demonstrate substantial variation in housing affordability across metropolitan areas, with counties in the Pittsburgh region generally showing more favorable affordability ratios than those in Philadelphia and Harrisburg. Forecast models predict continued pressure on affordability in most counties through 2028, though with notable differences in trajectory. This report provides stakeholders with

data-driven insights to inform housing market decisions, policy development, and investment strategies across Pennsylvania's diverse regional economies.

Introduction

In recent years, the housing market in the United States has had a noticeable price increase. Houses have become less affordable than ever, leading to critical effects on the economic stability of various regions of the United States. This report focuses on Pennsylvania's major metropolitan areas, including Philadelphia, Pittsburgh, and Harrisburg, and the surrounding counties comprising these regions. The aim is to examine the dynamics of housing affordability across these metropolitan areas, identifying trends and forecasting future conditions. This will provide young home buyers and government officials with valuable insights into the various metropolitan areas of Pennsylvania.

For this analysis, we leveraged two datasets to construct a comprehensive dataset. The first dataset, Redfin's housing market data, contains time series data on housing prices, supply, and market behavior while covering metropolitan areas with granular county-level statistics. The second dataset is derived from the American Community Survey from the U.S. Census Bureau, which offers a vast collection of economic indicators. By merging the two datasets, which span from 2012 to 2023, we combined the key features from both sources but also created additional composite features to enhance our analysis. It is important to note that data from 2024 was excluded, as the U.S. Census Bureau has yet to release the 2024 American Community Survey results.

Through our research, two modeling approaches were considered: time series forecasting using ARIMA (Autoregressive Integrated Moving Average) and machine learning prediction using XGBoost. Using 2 different models, we were able to capture different facets of the data. The ARIMA model accounted for the temporal patterns in the housing markets, while the XGBoost model accounted for complex relationships between affordability and socioeconomic factors. To obtain an affordability metric, we

created an affordability ratio, which is calculated as the median sale price divided by the median household income.

Several key questions emerged throughout our analysis:
What trends are present in housing costs and income levels across Pennsylvania, and do these trends affect affordability?

What are the differences in housing affordability between the different main metropolitan areas of Pennsylvania, and which counties offer the most favorable conditions for homebuyers?

What insights do our findings provide regarding the urgency for targeted affordable housing initiatives in certain regions?

Which areas are poised for growth in the upcoming years, and which areas are already well-developed with potentially limited opportunities for new investment?

By tracking affordability metrics and forecasting their trajectory, our comprehensive analysis provides home buyers, government officials, and investors with clear data-driven insights. These forecasts not only reveal current market conditions but also highlight long-term trends that can influence policy decisions and investment strategies. Overall, this will guide stakeholders in navigating the Pennsylvania housing market within the coming years.

Methods
Our analysis was devised using our two principal datasets as part of a two-sided analysis strategy to investigate and forecast housing affordability in Pennsylvania's major metropolitan areas. The sections below detail the data acquisition and preprocessing steps, the exploratory data analysis that guided our approach, and the modeling techniques that were used along the way.

Data Acquisition and Preprocessing

Our analysis focused on two main data sources: Redfin's housing market data and socioeconomic indicators from the American Community Survey (ACS). These datasets were merged to form a comprehensive time series dataset that captures both housing market behavior as well as the economic conditions across Pennsylvania from 2012 to 2023.

To create the dataset, ACS variables were first transformed from long form into wide format. The ACS variables are also reported annually, while the housing market data is reported monthly. Initially, we performed linear interpolation to estimate the monthly values from the yearly ACS variables. However, this presented issues with portions of the data at the ends being flat, so instead, we decided to use the yearly data as it is, as the economic factors of the regions will not be massively shifting from month to month. We then merged the socioeconomic indicators with the more granular Redfin data using shared identifiers such as county names, state, and the time period. Additionally, composite features were constructed to enhance the dataset, including percentage metrics for home ownership, education, and other socioeconomic data.

Exploratory Data Analysis

Before proceeding with any models, we performed extensive exploratory analysis to uncover some of the trends in the data and check the validity of the merged dataset that we had created. Time series visualizations were generated to display the trends in median sale price, inventory, and market supply for the overall market and for the Pennsylvania Metropolitan areas.

Originally, the idea was to evaluate the affordability ratio for all of the counties of Pennsylvania, although after we ran some initial visualizations, we ran into some issues. The main one being that the smaller county housing markets across Pennsylvania were highly variable and would present massive difficulties for

prediction. This is seen in the plot of year-over-year growth across all Pennsylvania counties (Figure D). The average growth for a county was 7%; however, there were multiple outliers, with the most significant being Forest County with -1.14%, McKean County with 1.89%, and Cameron County with 12.75%. There was also a lack of observations for smaller counties with many having less than 30 housing transactions per year, leading to massive differences in data with no apparent patterns.

The visualizations highlighted both long-term trends and more volatile periods which are similar to our current housing market. The main long-term trend was a power dynamic between sellers and buyers. In the summer months, the housing market is most active. This allows sellers to have more power as there is high demand from buyers looking to purchase new homes. This drives up the prices, active listings, and the percentage of price drops from the initial listing (Figures A, B, and C in the appendix). Then, at the starts and ends of the year, the prices, active listings, and percentage of price drops all shrink as the market cools off.

Given these strong historical trends in the data, as well as the more volatile periods, it makes the most sense to pursue an ARIMA model, which will be poised to leverage the historical data and an XGBoost model that will be able to adapt to unexpected market fluctuations given the volatile status of the current market. Then we will compare and contrast the two models to decide which is more suitable to use for the predictions out to 2028.

ARIMA Versus XGBoost Model Analysis

We first fit the models using data from 2012 to 2020, then used them to forecast out to 2023. While the ARIMA model effectively captured historical temporal patterns, it struggled with adapting to unpredictable changes, especially when socioeconomic factors played a larger role in the housing market. This limitation led us to explore machine learning techniques, and we ultimately chose the popular and versatile XGBoost model. We trained and fine-tuned the XGBoost model using its built-in

feature selection capabilities along with a grid search approach. In our final comparison (Figures E and F), XGBoost demonstrated better performance, as shown by lower mean absolute error and lower residual mean squared error.

Based on these results, we used the optimized XGBoost model to forecast housing affordability trends out to 2028, leveraging our full dataset from 2012 to 2023. The 2028 predictions (Figure G) look reasonable and reflect a likely market trajectory. Additionally, the variable importance analysis (Figure H) reveals that the top three predictors are median price per square foot, the percentage of the population with a bachelor's degree, and the total population in the area. This shows that the model is also properly leveraging the socioeconomic and housing market features to get well-rounded predictions for the markets.

Economic Disparities in Housing Affordability

To answer our research question of what the current housing market's trends are, we visualized the trends between median household income and median housing price (Figures I and J). The results revealed that for many of the metropolitan areas of Pennsylvania, housing prices are increasing faster than incomes, making home ownership increasingly difficult. The widening gap highlights a growing economic disparity across Pennsylvania. The only counties that are defying this trend are Cumberland, Westmoreland, and Philadelphia County, which have seen income growth keeping pace with housing prices. This means that there are still some relatively affordable counties, but the overall trend is a decline in affordability for most counties.

Affordability Forecasts and Comparisons for Pennsylvania Counties

To answer the research question: What are the differences in housing affordability between the different main metropolitan areas of Pennsylvania, and which counties offer the most favorable conditions for homebuyers? We first visualized the culmination of the 2028 projections for affordability for each county into their respective metropolitan areas (Figure K). Between the predicted metropolitan areas of

Harrisburg, Pittsburgh, and Philadelphia, the area that came out with the best affordability ratio was the Pittsburgh metropolitan area, with a predicted affordability ratio of 2.76. This was closely followed by the Harrisburg metropolitan area, which came in with a predicted affordability ratio of 3.18, and then the Philadelphia metropolitan area with a predicted affordability ratio of 3.76.

Then, breaking down the individual results of the counties that make up the metropolitan areas and visualising it (Figure L), the predicted most affordable counties for each metropolitan area were Westmoreland County for the Pittsburgh metropolitan area, with a predicted affordability of 2.5. Then, the predicted most affordable county for the Harrisburg metropolitan area was Perry County, with a predicted affordability of 2.07. Finally, the predicted most affordable county for the Philadelphia metropolitan area was Delaware County, with a projected affordability ratio of 3.21.

Governmental Housing Affordability Initiative Recommendations

Next, we sought to answer our research question of which areas might need governmental intervention in the form of targeted affordable housing initiatives. To answer this, we culminated our results into two main figures, these being Figures M and N. In recent years, Pennsylvania lawmakers have been floating the idea of trying to get all counties to follow a 3 times rule for housing affordability. This is unofficial as of now, but it would mean that the optimal affordability ratio of a county should be around 3, meaning that the price of the median house should be roughly three times more than the median household income. Based on our results (Figures M and N), it is evident that some of the counties with large affordability ratios may need some assistance through targeted affordability initiatives to lower the affordability ratio towards the government's target of 3. This especially applies to counties that already have a high affordability ratio and are forecasted to keep rising. The worst case of this is Chester County, the richest county in Pennsylvania, although these recommendations should also be considered in relation to the social factors of the region. Chester County specifically is unique as there are many mansions and some of

the best schools in the state, so getting a house there is seen as more of an investment than in other counties, which could be part of its high affordability ratio. Though if we were to make a general list of possible areas for targeted affordable housing initiatives, it would be the following counties and their respective predicted affordability ratios: Chester County with 4.07, Bucks County with 3.96, Philadelphia County with 3.84, and Montgomery County with 3.61. It must also be mentioned that a low affordability ratio could hint at economic troubles as well, with the general implication being that the area is economically depressed or stagnating and may require more investment. The only concern in this category would be Westmoreland County, with a ratio of 2.52 that is projected to be decreasing.

Conclusion

Based on our comprehensive analysis of housing affordability across Pennsylvania's metropolitan areas, several key conclusions can be drawn that provide valuable insights for stakeholders looking to navigate the future market.

Key Findings

The dual modeling approach we implemented using ARIMA and XGBoost has revealed significant variations in housing affordability across metropolitan areas. The Pittsburgh region generally demonstrates more favorable affordability ratios compared to Philadelphia and Harrisburg counties, with an overall predicted affordability ratio of 2.76, followed by Harrisburg at 3.18 and Philadelphia at 3.76. This suggests that prospective homebuyers may find more accessible housing options in specific counties rather than broadly across entire metropolitan regions.

Our county-level analysis identified the most affordable options within each metro area: Westmoreland County (2.5) in the Pittsburgh region, Perry County (2.07) in the Harrisburg area, and Delaware County (3.21) in the Philadelphia metropolitan area. These counties represent potential opportunities for homebuyers seeking more affordable housing within their preferred region.

The affordability forecasts through 2028 indicate continuing pressure on housing affordability in most counties, with housing costs generally increasing faster than incomes. Philadelphia County, Chester County (4.07), and Bucks County (3.96) show particularly concerning trajectories, while counties like Westmoreland maintain relatively stable or even improving affordability metrics.

Feature importance analysis through XGBoost identified that median price per square foot, the percentage of the population with bachelor's degrees, and the total population are among the most influential factors affecting housing affordability. This highlights the complex interplay between demographic factors, educational outcomes, and housing market dynamics.

Limitations and Further Research

Despite the methodology employed, several limitations should be acknowledged. First, our models do not account for potential economic shocks or policy changes that could significantly alter housing market dynamics. The COVID-19 pandemic demonstrated how unexpected events can rapidly transform housing preferences and pricing patterns, suggesting that our forecasts should be interpreted with appropriate caution.

Additionally, while our dual modeling approach provides complementary perspectives, with XGBoost demonstrating superior performance based on error metrics, further research could benefit from ensemble methods that explicitly combine ARIMA and XGBoost predictions to improve forecast accuracy. Incorporating additional variables could also enhance the models' predictive capabilities.

Future research should explore more geographical analysis, potentially at the zip code or census tract level, to capture neighborhood-specific affordability trends that county-level analysis may obscure. Gaining additional historical data could also improve the models' ability to capture longer-term cyclical patterns in housing markets.

Appendix:

Figure A:

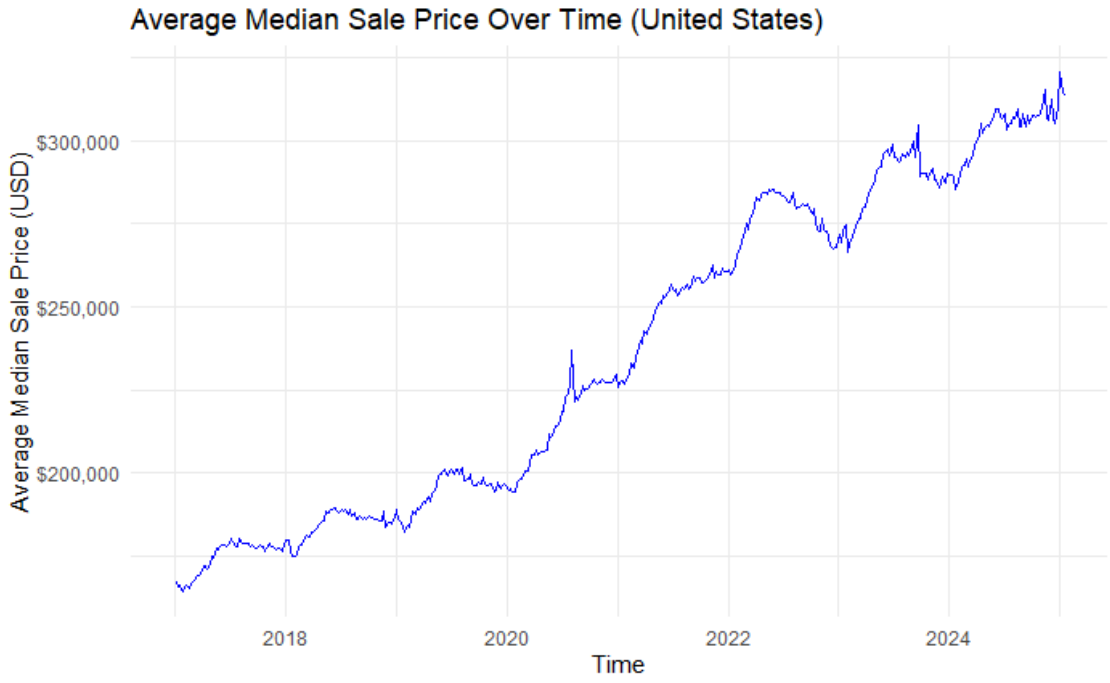

Figure B:

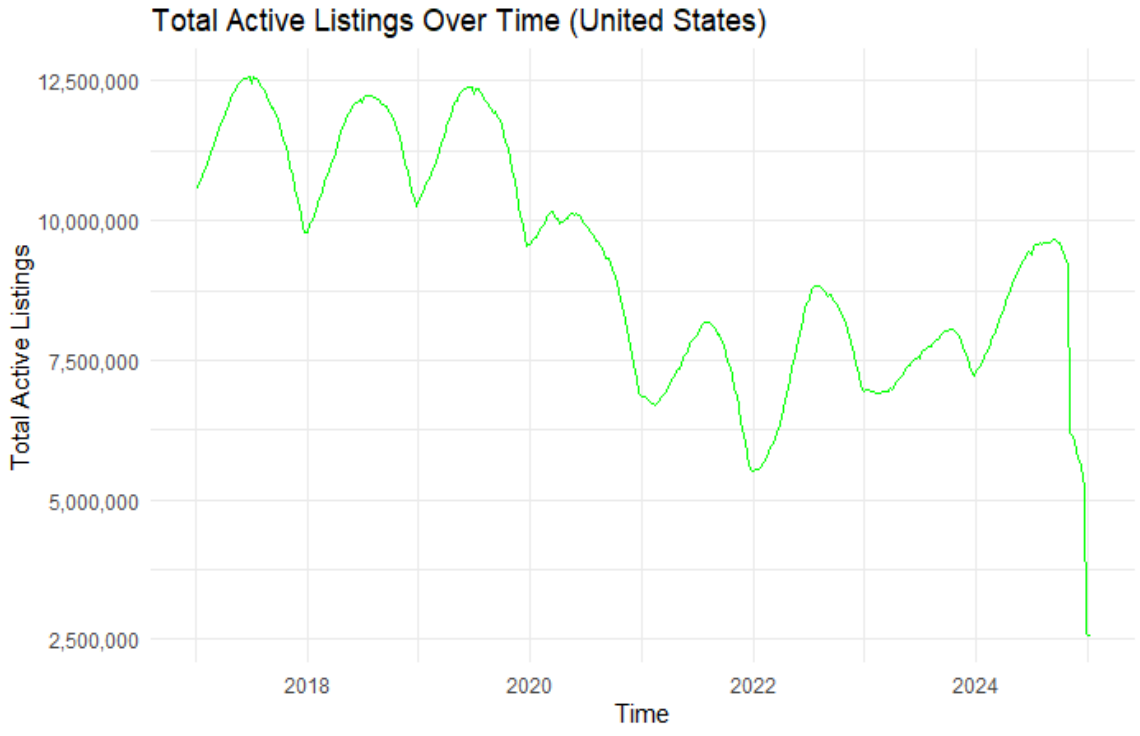

Figure C:

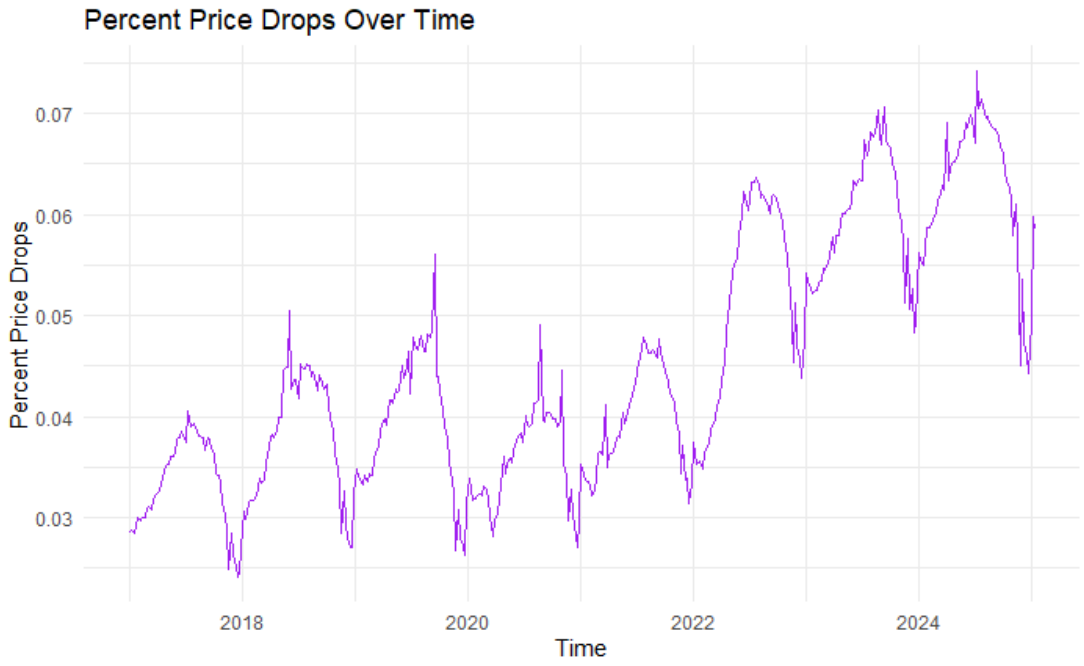

Figure D:

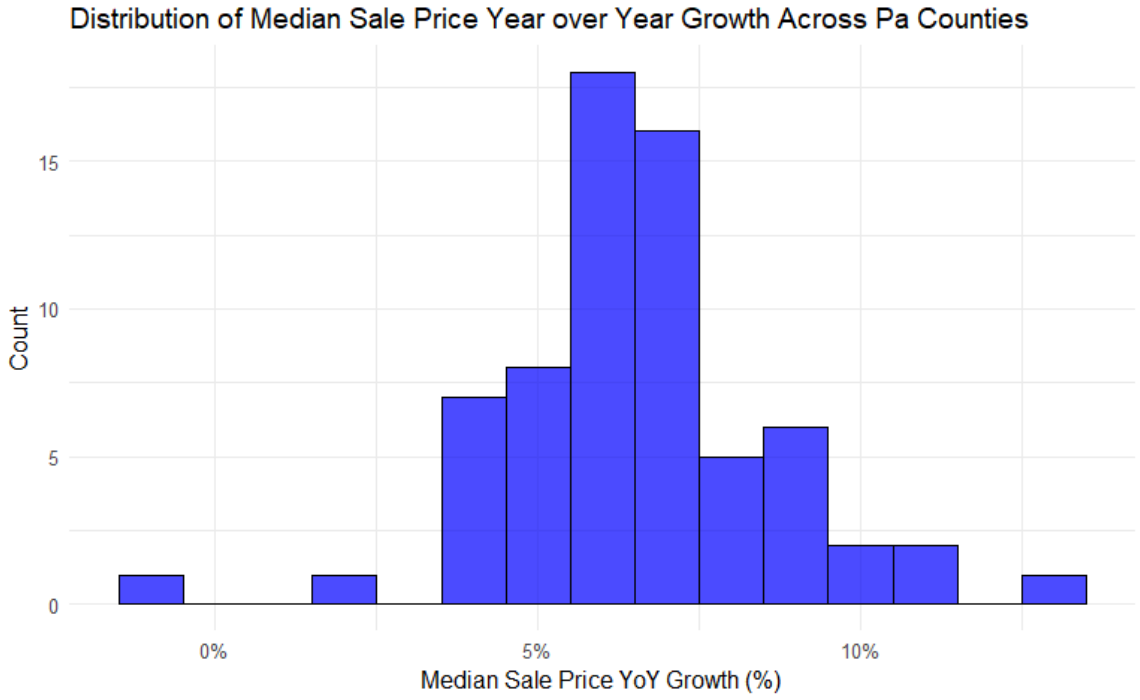

Figure E: The model metrics of the XGBoost model

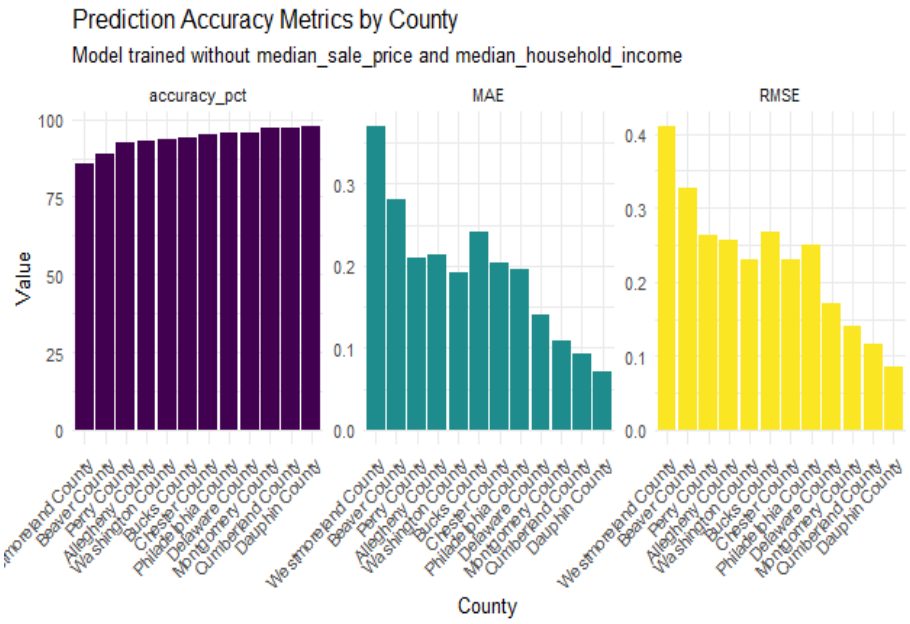

Figure F: The model metrics of the ARIMA model

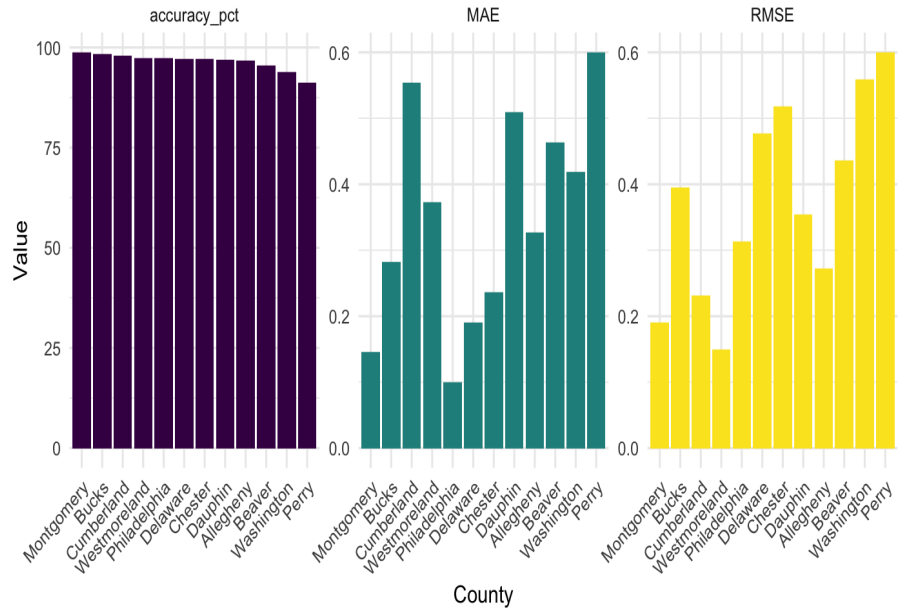

Figure G: The XGBoost model forecast to 2028

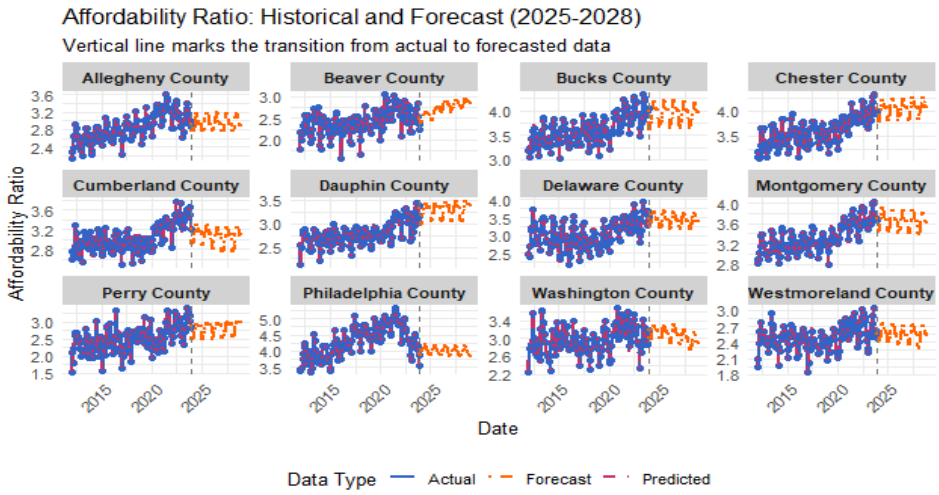

Figure H: The variable importance rankings of the top ten features for the 2028 forecast

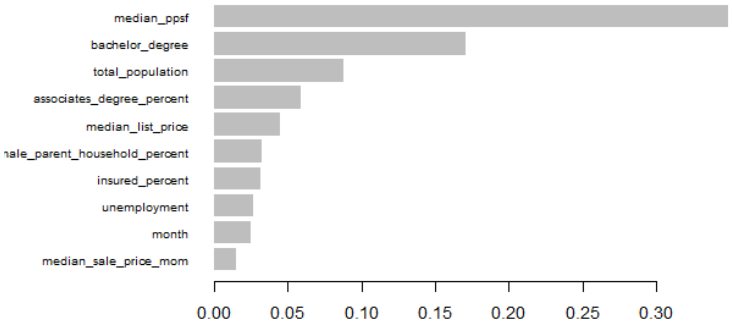

Figure I: Trends of Housing Prices VS Income

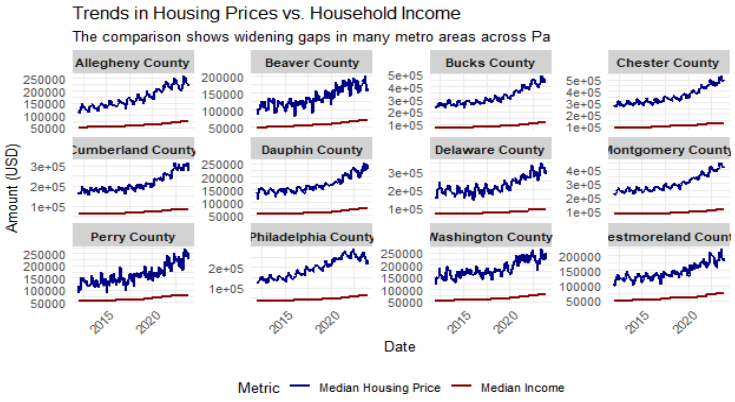

Figure J: Housing Price Growth VS Income Growth bar chart

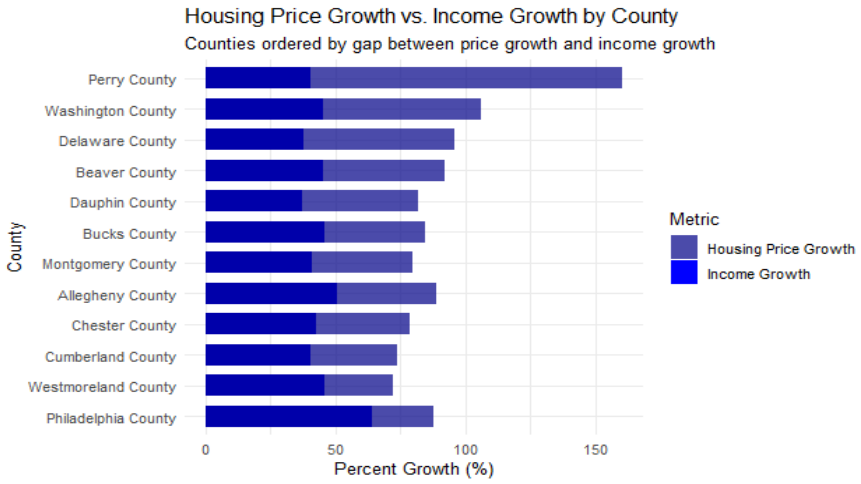

Figure K: The Metropolitan Areas Cumulative Affordability Forecast

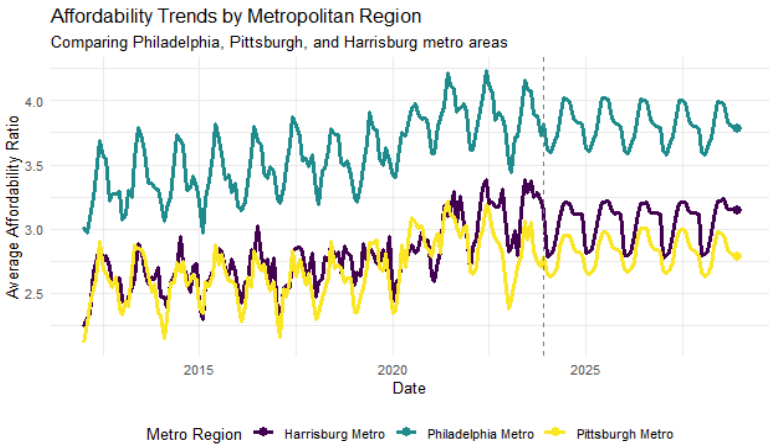

Figure L: Affordability Forecasts for the Metropolitan Counties

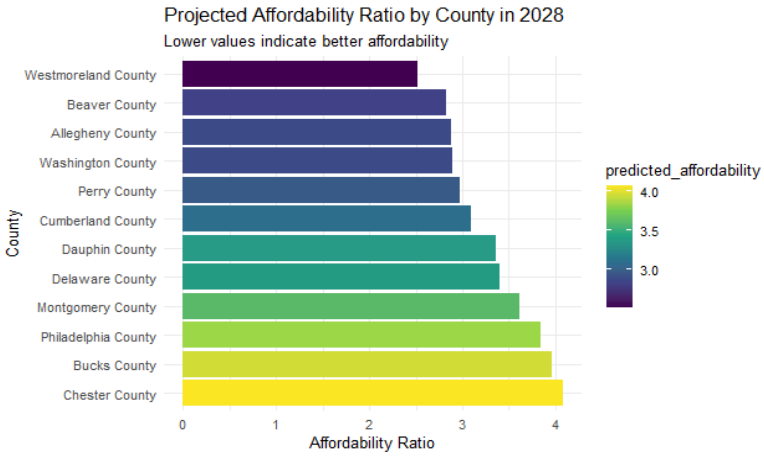

Figure M: Projected Change in Affordability

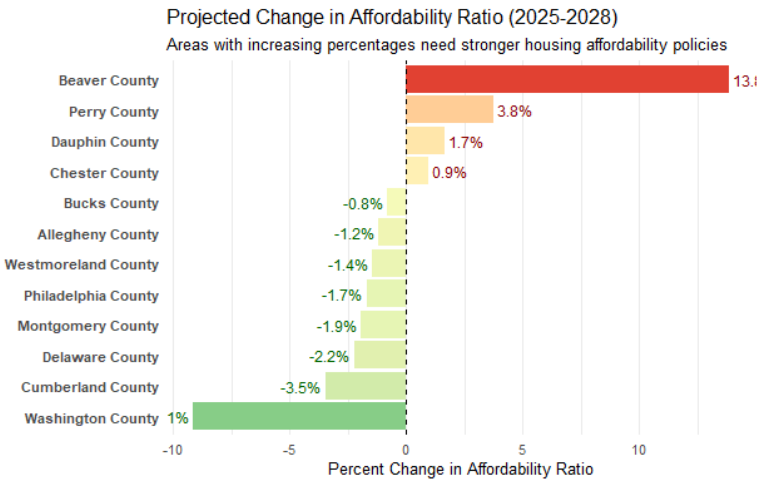

Figure N: Forecasted Affordability Ratio with Improvement Status