Unveiling Housing Market Dynamics

An Analysis of the U.S. Housing Market Trends and Insights (2017–2025)

By: Colton Dumm and Ryan Quinlan

Project Introduction

Objective: Understand trends and insights in U.S. housing data using Redfin's dataset.

Key Questions:

1. How have median sale prices changed over time?
2. What trends exist in housing supply and demand?
3. Are there relationships between supply and pricing?

Dataset Overview

Data Source: Redfin Housing Market Data

Key Features:

- Time series data on housing prices, supply, and market behavior.
- Covers metro and county-level statistics.

Exploration Steps: Structure, summary statistics, categorical data review.

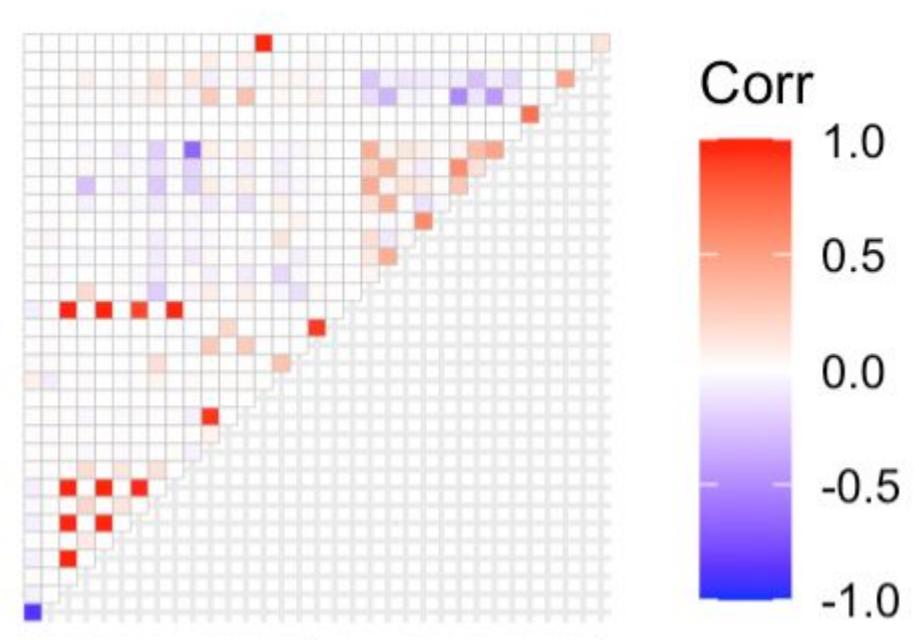
Part 1
Exploratory Data Analysis

Overview

Summary of the Types of Features in the Redfin Dataset

1. **Time Period Information**: Covers the start and end dates of the data period.
2. **Geographic Information**: Describes the regions included in the data (County, State)
3. **Listing and Sales Metrics**: Includes data on new listings, pending sales, and off-market listings.
4. **Sales and Price Metrics**: Focuses on prices and price per square foot.
5. **Market and Inventory Metrics**: Provides data on active listings, days on market, price drops, and inventory age.
6. **Sale Performance Metrics**: Relates to square footage of pending sales and sale-to-list ratios.

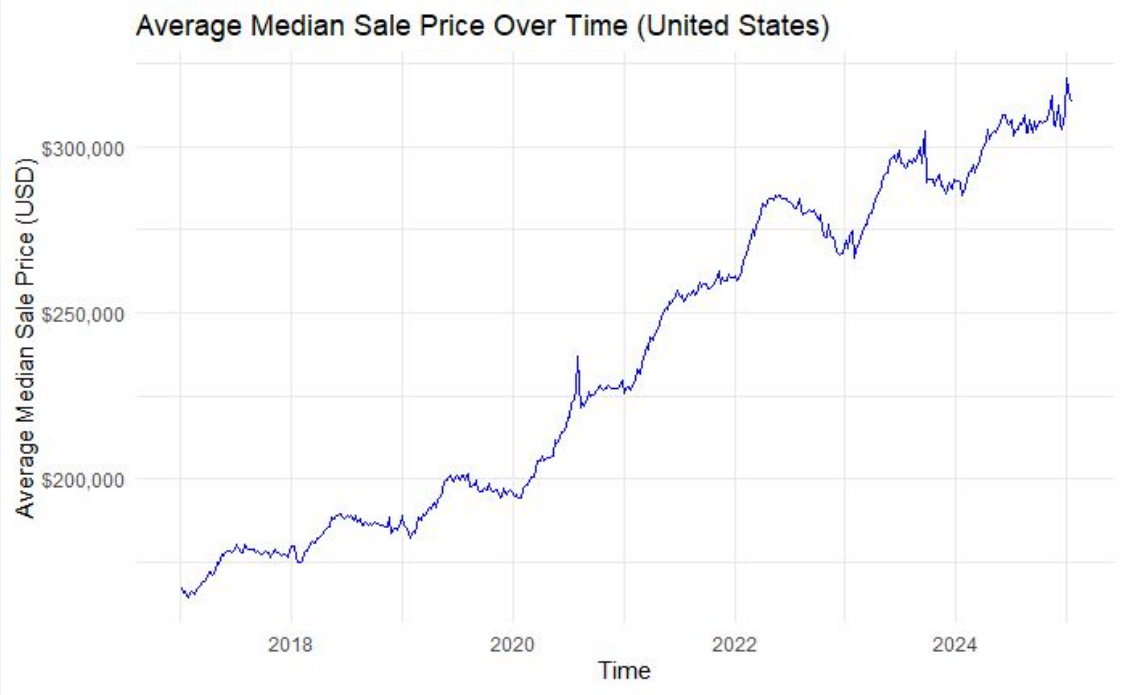
Addressing Correlation



Breakdown

- A correlation matrix helps us eliminate any multicollinearity between the featured variables
- Used a half correlation matrix that shows strength and direction of correlations
- Most of the variables that strongly correlate are very similar in name
- i.e. Region type ID and Region type: -0.89

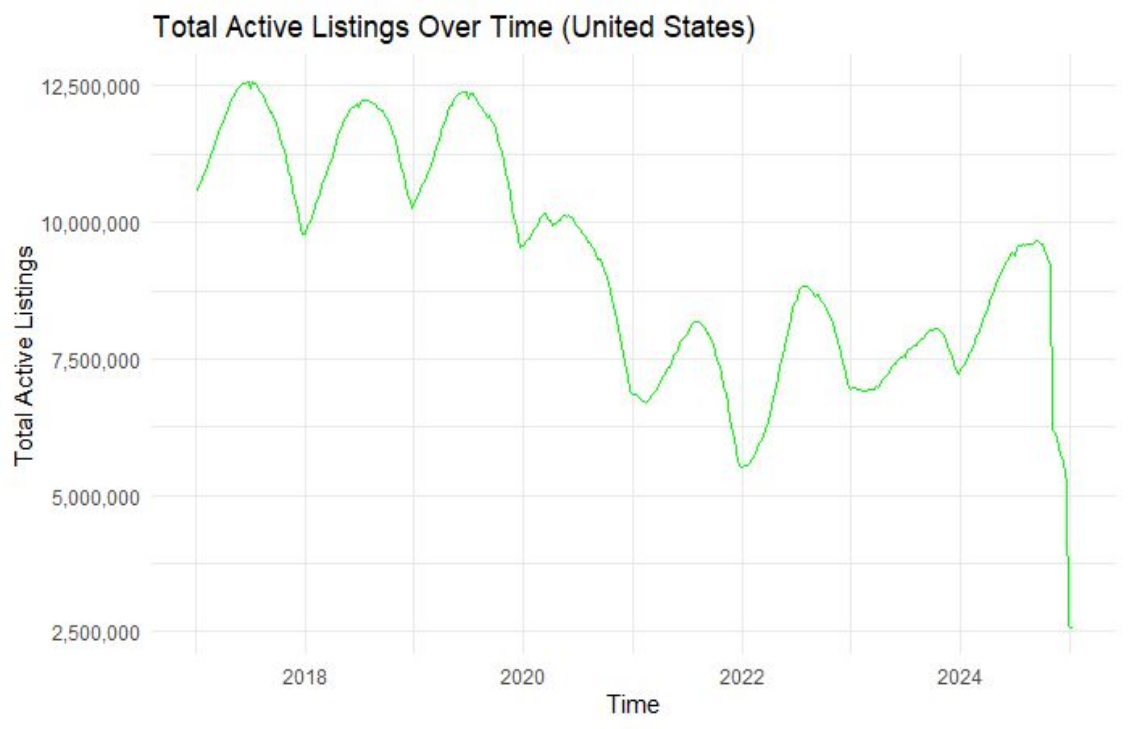
Average Median Home Price



Breakdown

- We can see that the U.S. housing market has been fairly stable for the last 8 years.
- There is a clear and steady rise in median home prices.
- However, there is one significant spike from 2020 to 2022, likely due to the market disruption
- There is seasonality which suggests housing prices vary within each year, due to real estate selling seasons

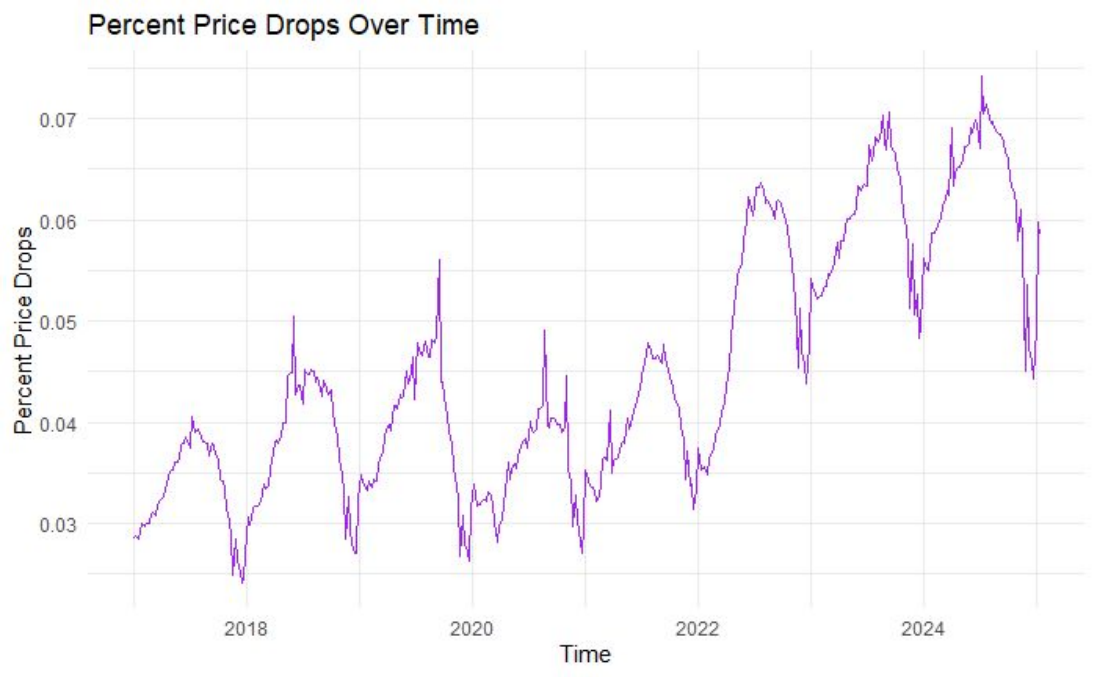
Active Listings



Breakdown

- **Cyclical Nature of Listings:**
  - **Low Listings:** Begin and end of the year.
  - **High Listings:** Peak during summer months.
- **Recent Decrease in Listings:**
  - **Possible Causes:**
    - Sellers may be hesitant due to market uncertainty.
    - Rising home prices may limit buyer accessibility.
    - Homeowners staying in properties longer due to high mortgage rates.

Price Drops on Active Listings



Breakdown

1. **Cyclical Patterns**

   - Summer: High inventory = Price drops to compete.
   - Year-End: Low inventory = Prices hold steady.

2. **Recent Shifts** (Price drops rising 2020–2024):

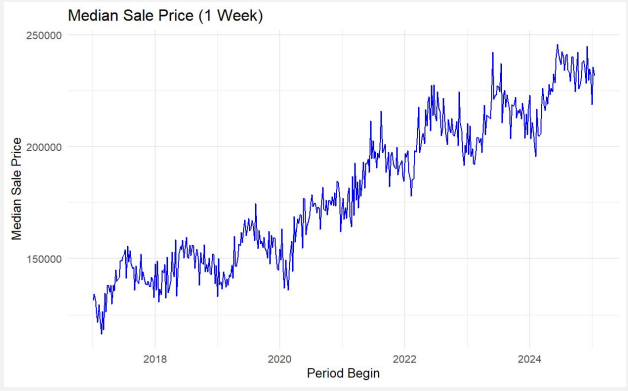
   - Overpriced listings: Sellers adjust to buyer resistance.
   - Buyer leverage: Fewer buyers demand steeper discounts.

Part 2
Research Questions and EDA cont.

Research Questions

1. What are the top 3-5 projected most affordable county and metro markets to live in the state of Pennsylvania?
2. Can we develop a market health score for each region based on multiple indicators, and predict future market performance?
    a. Combination of metrics to create a composite score and predict its movement.
3. Is it possible to determine if a county is a buyers, sellers, or balanced market?
4. Can we predict the length of time a house will be on the market?
    a. We could also break down the factors that allow houses to sell in shorter amounts of time.

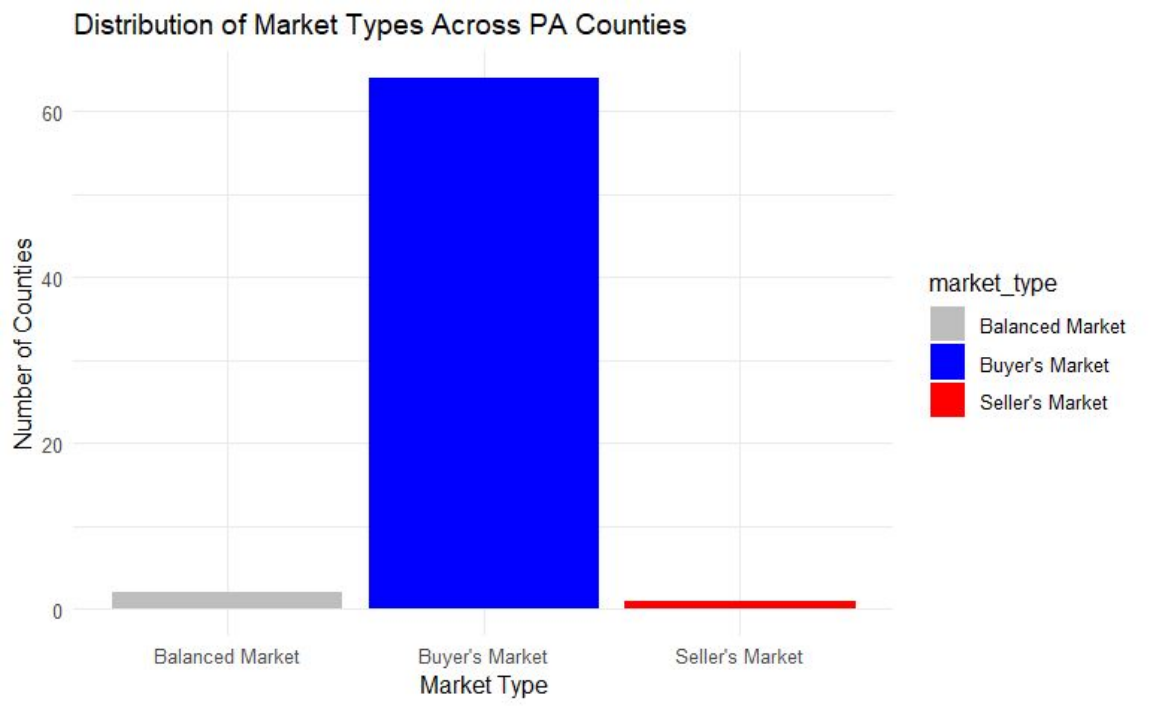
Median Sale Prices Across Time Frames



Visualizing median home prices through three distinct observation windows:

1. **Weekly Observations**
   - Captures short-term fluctuations and immediate market dynamics.
   - Useful for identifying rapid changes or anomalies in pricing trends.
2. **Monthly Observations**
   - Smooths out weekly volatility to reveal broader monthly patterns.
   - Highlights seasonal trends and more stable price movements over time.
3. **Quarterly Observations (Our Primary Data)**
   - Provides the most comprehensive view by aggregating data over three months.
   - Ideal for analyzing long-term trends and sustained shifts in the housing market.

Market Types Across PA Counties



Buyer's Markets

- Definition : >7 months of housing supply.
- Characteristics : Favor buyers—more options, lower prices, greater negotiation power.

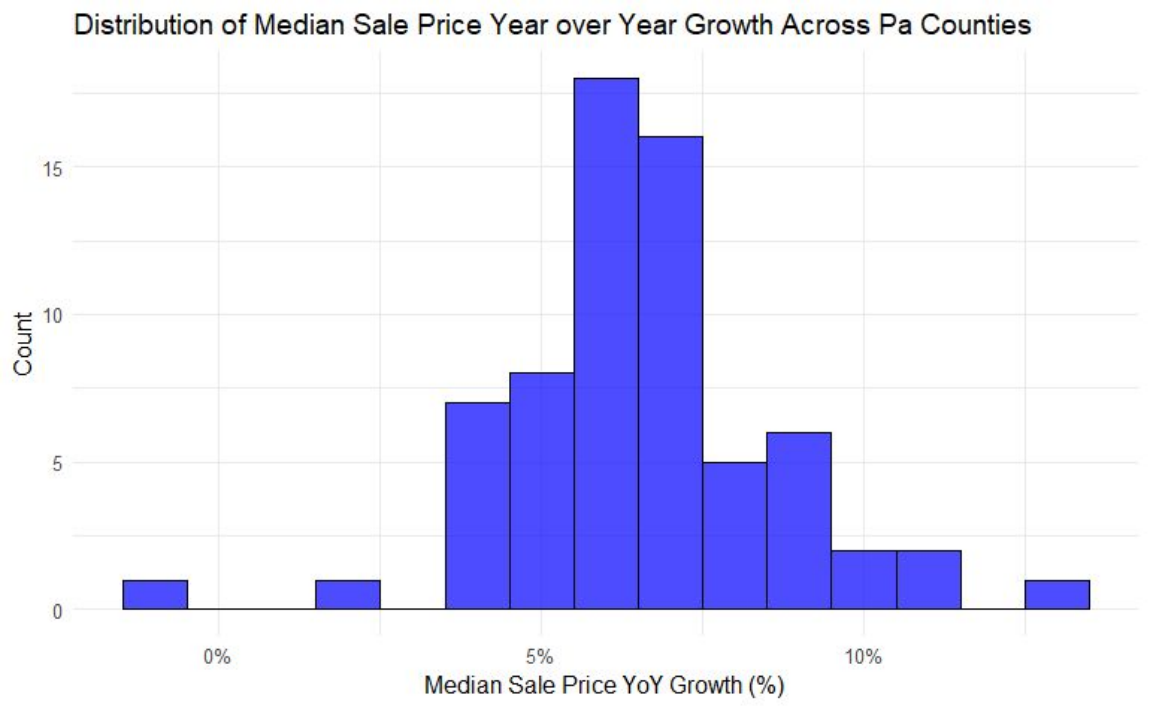
Seller's Markets

- Definition : <5 months of housing supply.
- Characteristics : Favor sellers—high demand, limited inventory, faster sales.

Balanced Markets

- Definition : 5–7 months of housing supply.
- Characteristics : Equal footing for buyers and sellers—stable pricing, balanced power.

Sale Price Growth by County



Most Counties of PA

- Grow roughly 7% year over year in Pennsylvania since 2017

The Upper Outlier

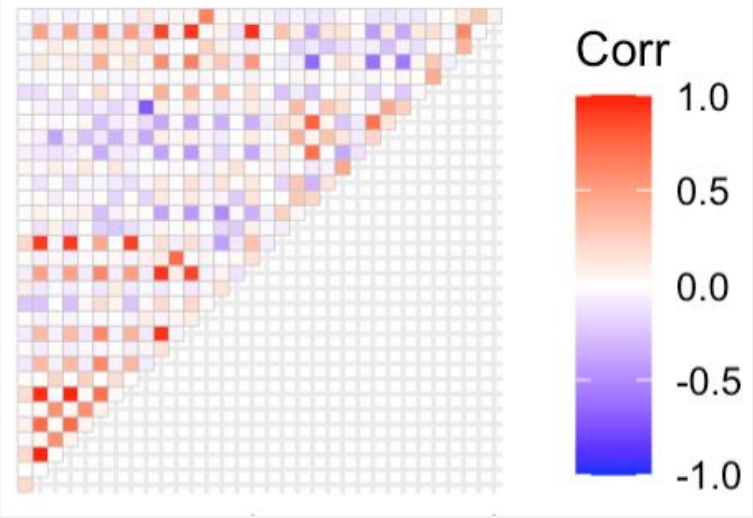
- Cameron County: 12.75%

The Lower End Outliers

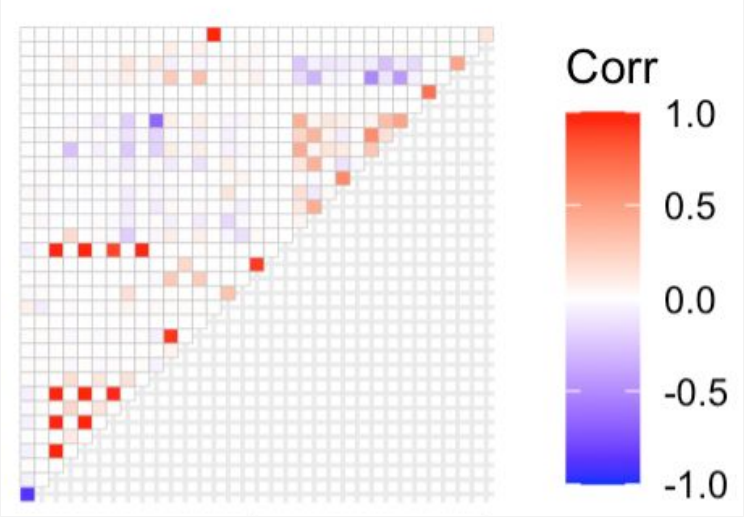
- Forest County: -1.14%
- McKean County: 1.89%

PA Correlation Matrix

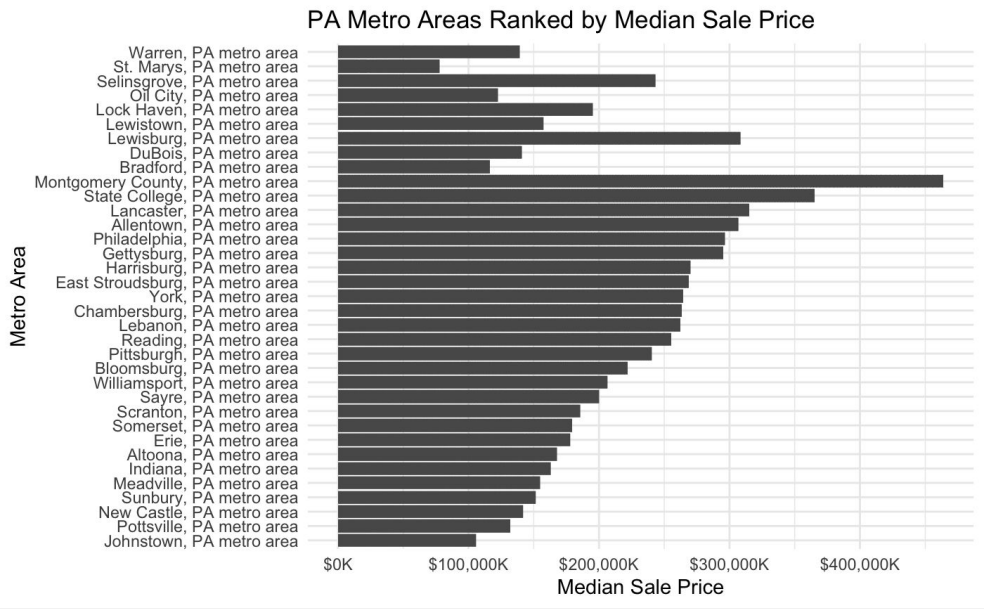
- Addressing multicollinearity may be needed upon further review
- Far more variables are correlated after focusing on specifically PA
- Reducing variables could make models more interpretable



- Old matrix has far less correlation between variables
- Hence why it was important to reevaluate this metric for PA
- May cause noise and overfitting in a potentia model



Median Price Metro Areas



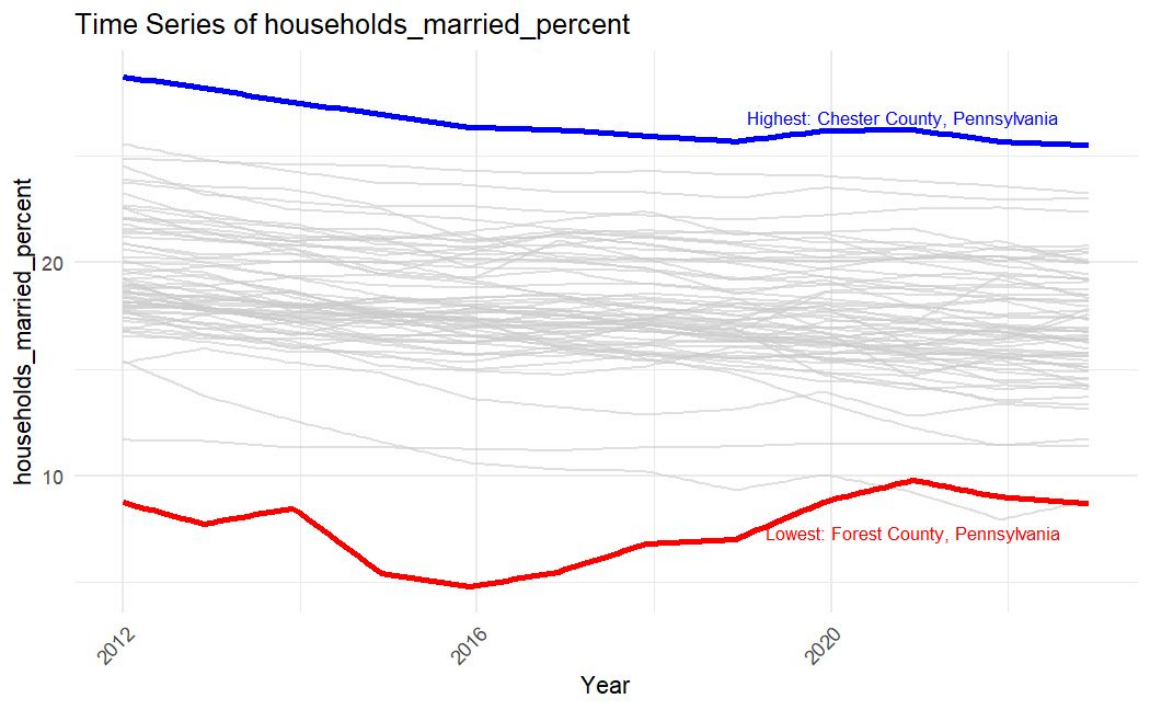
- Montgomery county, State College, Lancaster, Allentown, Philadelphia
- May focus on metro areas with a certain population threshold.
- Depends on research question
- Who is our audience? Buyers or Sellers?

Part 3
Research Questions, EDA, and Models

Research Questions

1. Using a 5 year forecast, what are the top 3-5 projected most affordable counties to live in the state of Pennsylvania?
2. Can we develop a market health score for each region based on multiple indicators, and predict future market performance?
    a. Combination of metrics to create a composite score and predict its movement.
3. Can we predict the length of time a house will be on the market?
    a. We could also break down the factors that allow houses to sell in shorter amounts of time.
4. Can we predict the rankings of each county by wealth and the predictors that are the leading factors to wealth accumulation in a county?

Married Couples Across PA Counties



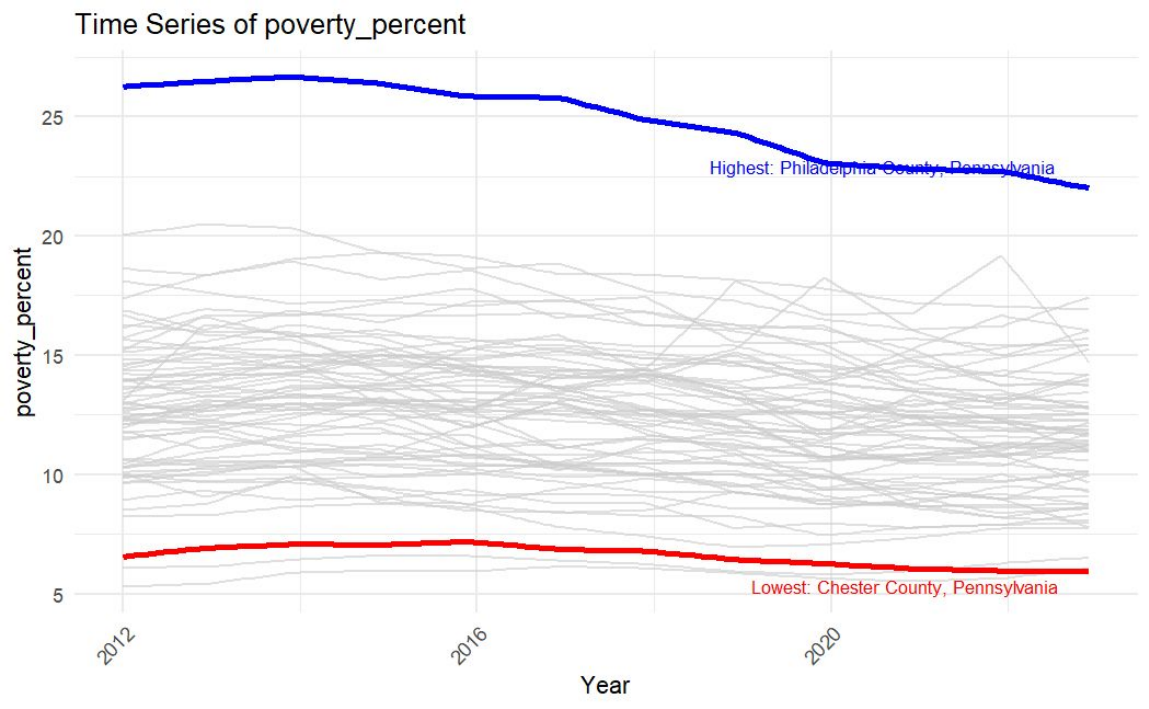
Implications:

- Chester County (the richest county in PA) has the best marriage rate by far whereas Forest County one of the poorest counties has the lowest marriage percent
- This could be an indicator of the effect of wealth on marriages and vice versa

The General Trend:

- Marriages seem to have decreased over the last decades this is likely due to two factors:
    - 1. A decrease in marriages
    - 2. A increase in divorces

Poverty Percent Across PA Counties



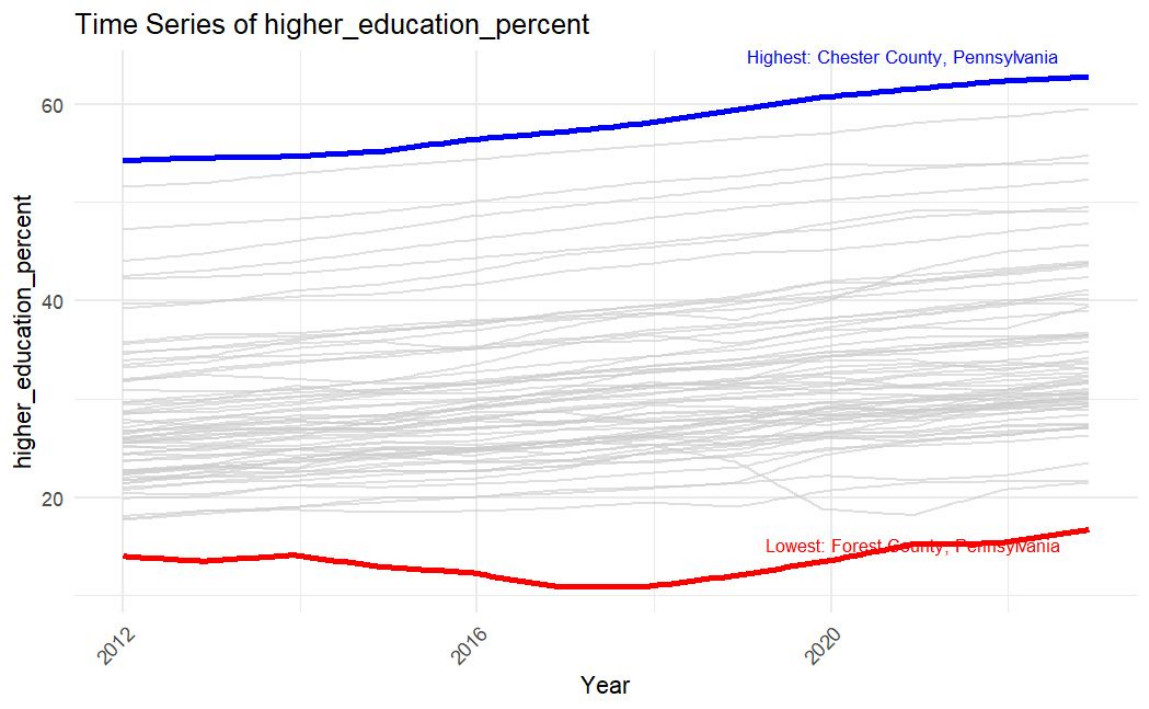
Implications:

- Chester County (the richest county in PA) has the lowest poverty rate by far whereas Philadelphia County one of the most densely populated countries has the highest poverty rate by far.

The General Trend:

- Poverty does seem to be decreasing across PA. This is great for the state and shows that we really are doing better than past decades in most counties.

Higher Education Across PA Counties



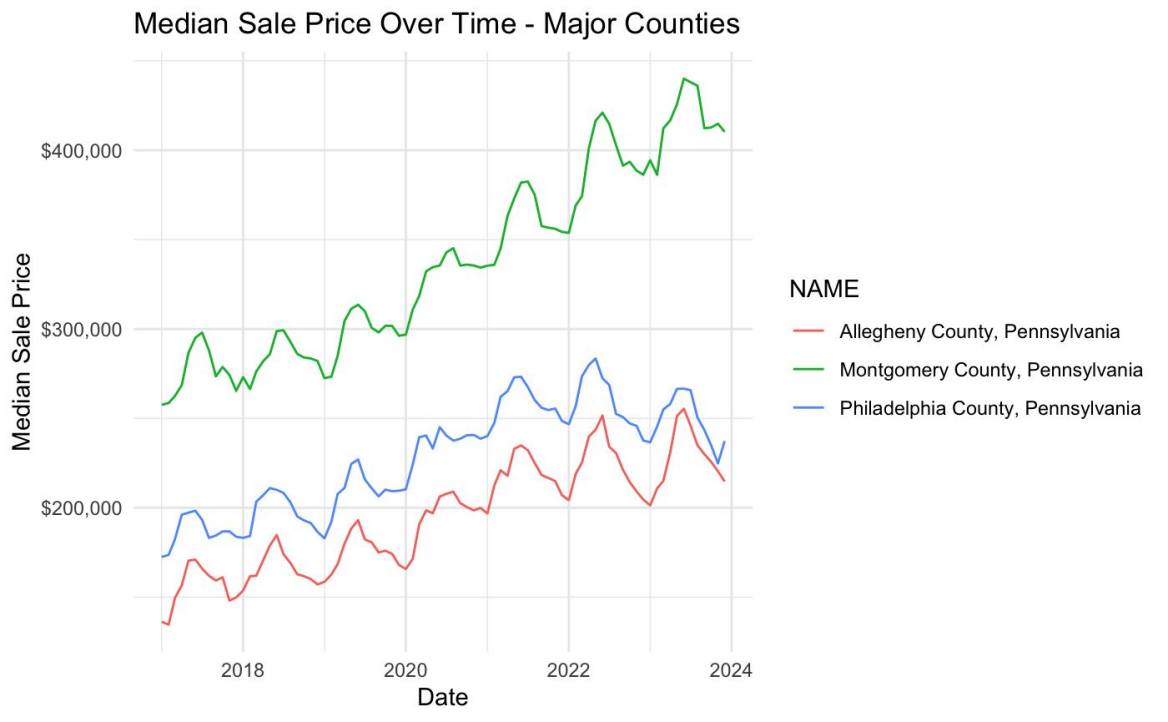
Implications:

- Chester County (the richest county in PA) has the best education rate by far whereas Forest County one of the poorest counties has the lowest higher education percent
- This could be an indicator of the effect of higher education on wealth and vice versa

The General Trend:

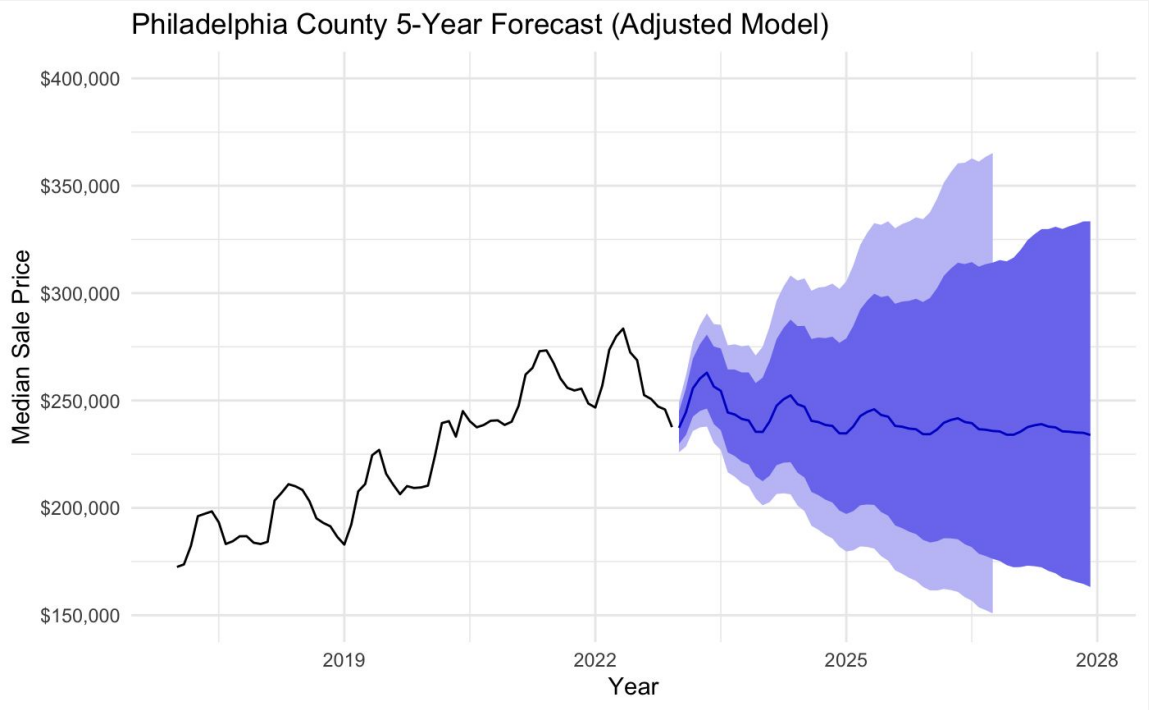
- The percent of people pursuing higher education is only increasing. This is great but also seem to be leading to the decrease in value of a higher education. Conversely this also means that trades are becoming more lucrative across PA.

Let’s look at 3 Counties



- All three counties show an overall upward trend from 2018 to 2024.
- There's a particularly steep price increase across all counties starting around 2020.
- **Montgomery County**
  - Has the highest jump in median price by nearly $150k from 20-24.
- Philadelphia and Allegheny counties appear to be experiencing more significant price corrections than Montgomery County.
- Seasonality present as expected

Philadelphia 5 year Forecast

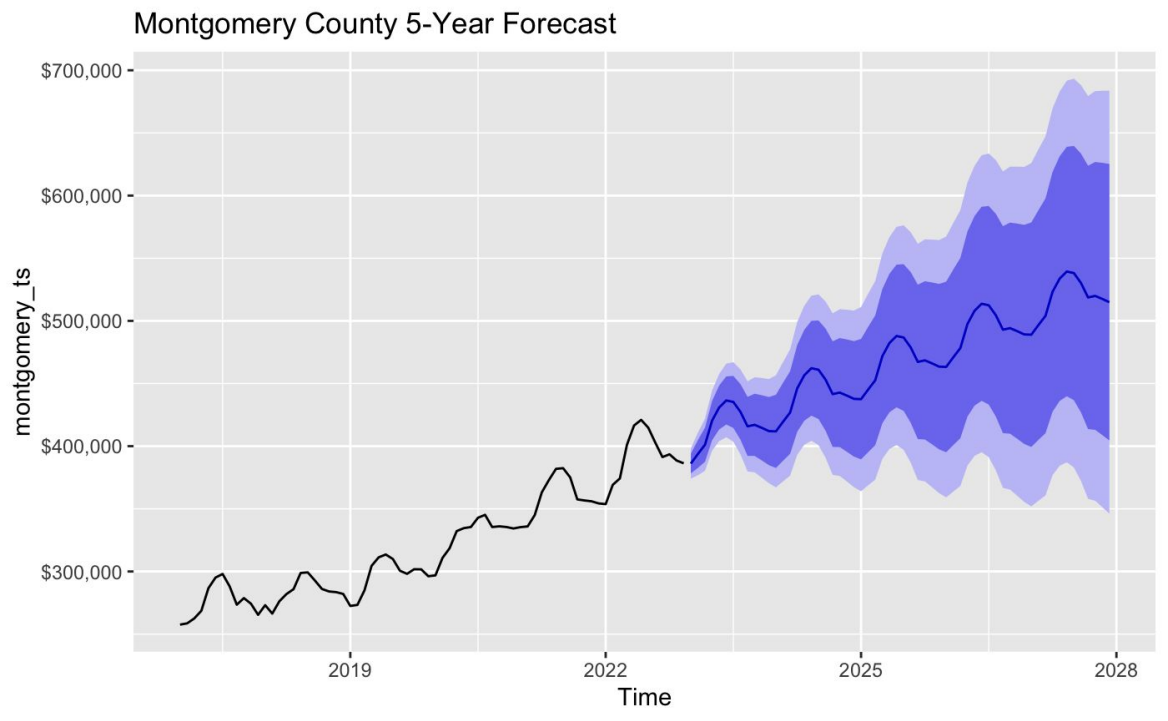


- The confidence intervals expand moderately over the 5-year forecast horizon, with an 80% prediction interval and 95% prediction interval reflecting increasing uncertainty over time while maintaining realistic bounds for housing market behavior.

Central Forecast

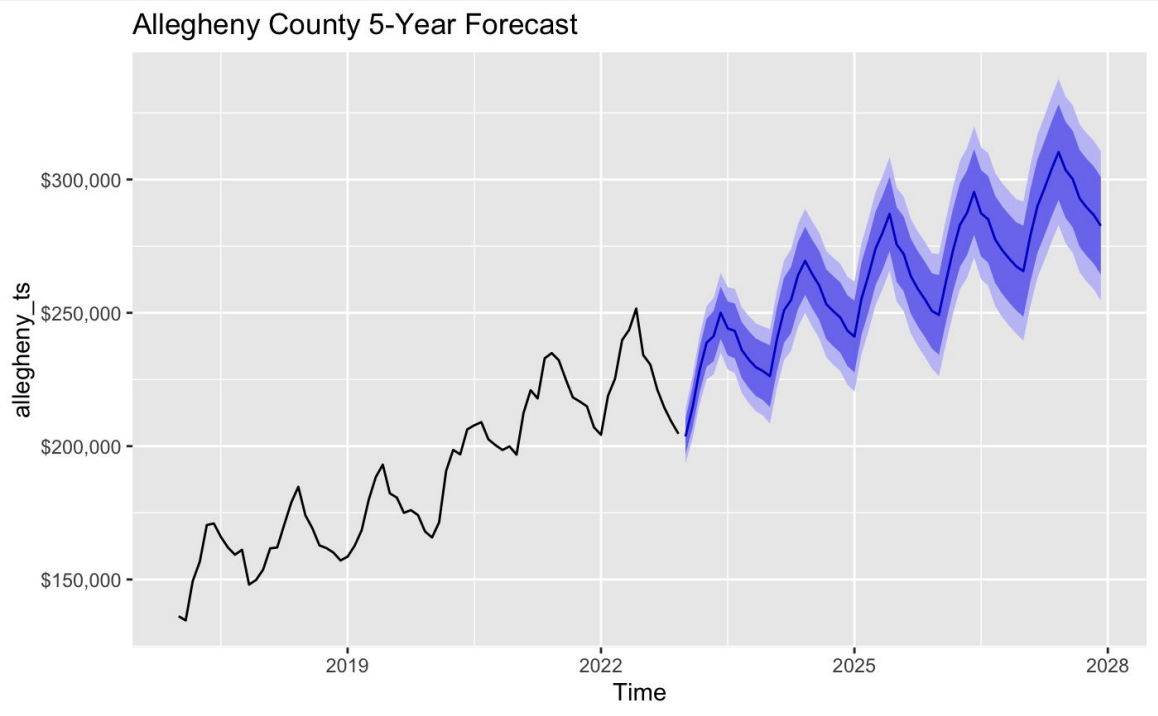
- Shows a slight downward trend from current levels
- Seasonality holds
- Used a box-cox transformation to stabilize the variance throughout the time series
- Toughest of the 3 counties to work with

Montgomery 5 year Forecast



- The confidence and prediction interval bands widen symmetrically
- The predicted forecast is expected to reach $550k median home value by 2028
- The upward trajectory is more confident than Philadelphia's forecast, suggesting stronger underlying market fundamentals
- Philly has a steadier forecast in comparison

Allegheny 5 year Forecast



- The confidence intervals are far more manageable with uncertainty over time in comparison
- Allegheny shows the most dramatic seasonal market fluctuation which continues in the forecast
- Strong upward trend that reaches $300k by 2028
- Easier market to predict as a whole

Part 4
Improvements and Further Results

Affordable Metro Counties

- We decided to focus on grouping the surrounding counties of the 3 prominent metropolitan areas of Pennsylvania
    - Harrisburg, Pittsburgh, and Philadelphia

- This will be used to compare an affordability metric quantified by creating an affordability ratio.
    - Median_sale_price / median_household_income
- The one limitation is that we only have annual household income data instead of monthly

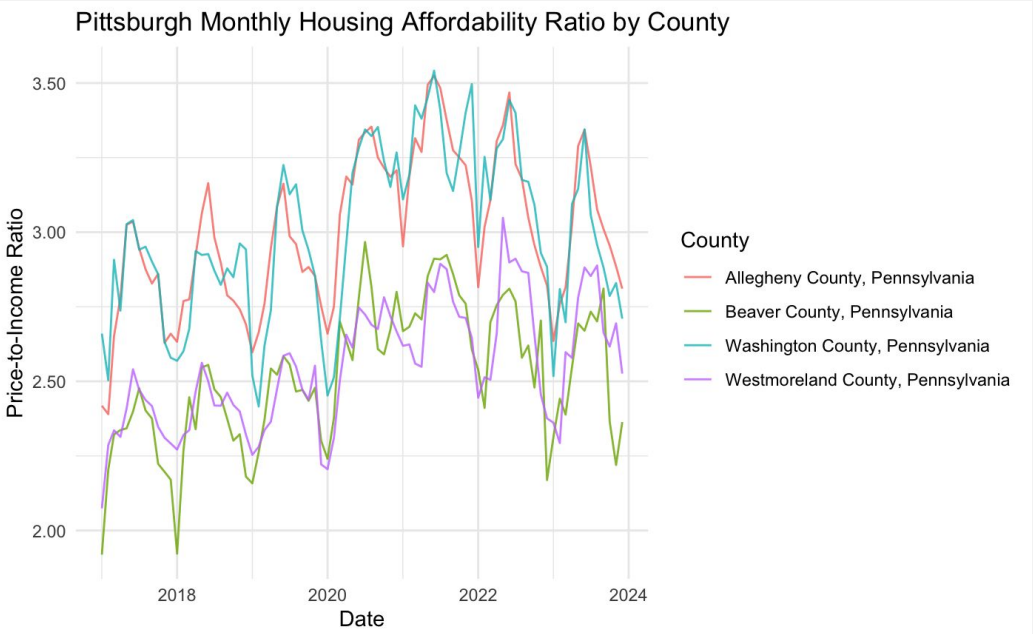
- Comparing within each county for those looking to buy houses in each metro area
- Forecasting future ratios to see whether they hold or not

- The goal is to provide buyers the most affordable county to live in for each metro area

Plots

- The monthly plots reveal how much affordability fluctuates seasonally and in response to market changes, which helps identify which counties have more stable or volatile housing markets.
- By calculating ratios rather than just looking at raw prices, you can directly compare affordability across counties with very different income and price levels.
- The affordability metric shows how many years of income would theoretically be needed to purchase a home outright.

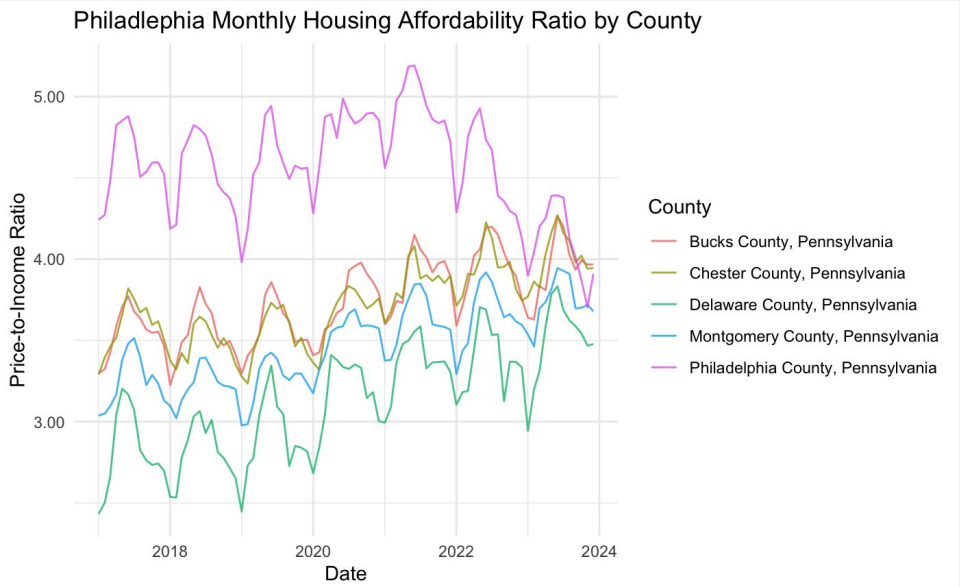
Pittsburgh Counties



NAME <chr>	median_household_income <dbl>	median_sale_price <dbl>	affordability_ratio <dbl>
Beaver County, Pennsylvania	60876.00	153717.8	2.525097
Westmoreland County, Pennsylvania	63438.14	162260.3	2.557771
Washington County, Pennsylvania	67224.86	201913.9	3.003561
Allegheny County, Pennsylvania	64809.71	195007.5	3.008924

- Pittsburgh has high volatility in the housing market
- The ratios are relatively low
- Washington County (turquoise line) shows the highest peaks, reaching above 3.5 at certain points
- Beaver County (green line) experiences the widest swings, from below 2.0 to above 3.0
- The graph shows regular cyclical patterns that suggest seasonal housing market fluctuations
- There's a general upward trend in ratios from 2017 to 2021, indicating declining affordability

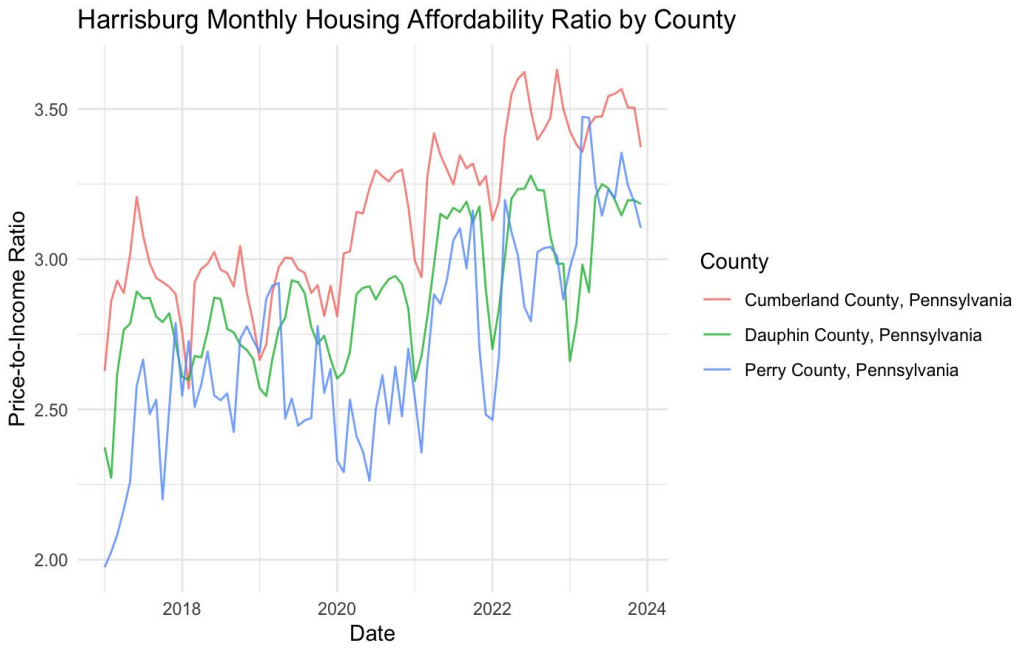
Philadelphia Counties



NAME <chr>	median_household_income <dbl>	median_sale_price <dbl>	affordability_ratio <dbl>
Delaware County, Pennsylvania	78208.14	246982.6	3.158016
Montgomery County, Pennsylvania	96620.57	336572.6	3.483447
Chester County, Pennsylvania	106443.14	397692.2	3.736194
Bucks County, Pennsylvania	95640.71	359339.4	3.757181
Philadelphia County, Pennsylvania	50047.29	228039.1	4.556473

- Again, most counties have the upward trend from 2017-2021.
- The most affordable county in Philadelphia is less affordable than every county in Pittsburgh.
- All of the counties start to converge towards the end of 2023.
- Philadelphia is by far the most surprising county because of the sharp downward trend from 2021-2023.
- Seasonality is present as expected.

Harrisburg Counties



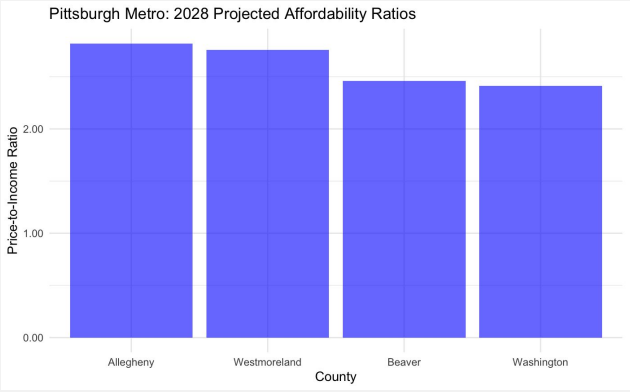
NAME <chr>	median_household_income <dbl>	median_sale_price <dbl>	affordability_ratio <dbl>
Perry County, Pennsylvania	69334.29	189986.0	2.740145
Dauphin County, Pennsylvania	64501.43	186904.4	2.897678
Cumberland County, Pennsylvania	74738.71	237051.2	3.171732

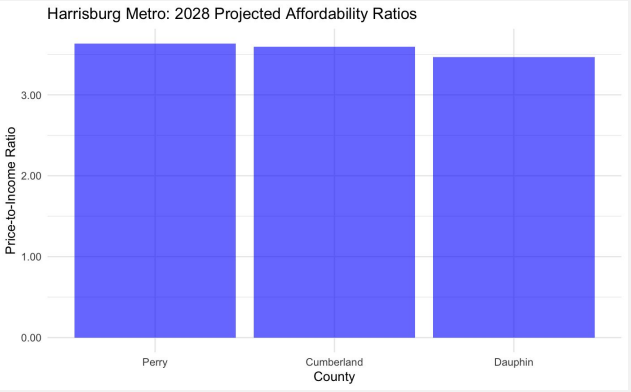
- The trend differs from the other metro areas as the upward trend does not end at 2022 and continues up to 2024.
- Less extreme volatility than Philadelphia metro
- Similar cyclical patterns to Pittsburgh, but with less dramatic swings
- There is a convergence similar to Philadelphia's at the end of the timeline.

Compound Annual Growth Rate

- Given the median_household_income limitation, we found that the best approach for now would be predicting the the compound annual growth rate for the same duration as our forecasted median_sale_price
- Income typically follows more stable long-term trends than housing prices, making a growth rate approach reasonable.
- CAGR provides an easily interpretable growth metric that captures the essential trend without overfitting to short-term fluctuations.

Settling on Predicted Ratios





- Forecasted each county and their median_sale_price using an auto.arima function
- To settle on the final ratios, we used the 2028 average forecasted median_sale_price and the predicted CAGR to obtain the median_household_income.
- Looking to improve this going forward by experimenting

- We can do this by trying different ARIMA parameters or seasonal components because auto.arima may not capture everything best.

$$\mathrm{CAGR} = \left(\frac{V_{\text{final}}}{V_{\text{begin}}} \right)^{1/t} - 1$$

```
# Now let's predict the household incomes for future years to be able to find updated affordability ratios
```{r}
project_income_cagr <- function(data, county_name) {
  # Get county data
  county_data <- data[data$NAME == county_name, ]

  # Get first and last year values
  start_year <- min(county_data$year)
  end_year <- max(county_data$year)
  start_value <- county_data$median_household_income[county_data$year == start_year][1]
  end_value <- county_data$median_household_income[county_data$year == end_year][1]

  # Calculate CAGR
  n_years <- end_year - start_year
  cagr <- (end_value/start_value)^(1/n_years) - 1

  # Project forward 5 years
  last_value <- end_value
  future_values <- data.frame(
    year = 2024:2028,
    projected_income = last_value * (1 + cagr)^(1:5)
  )
```
```

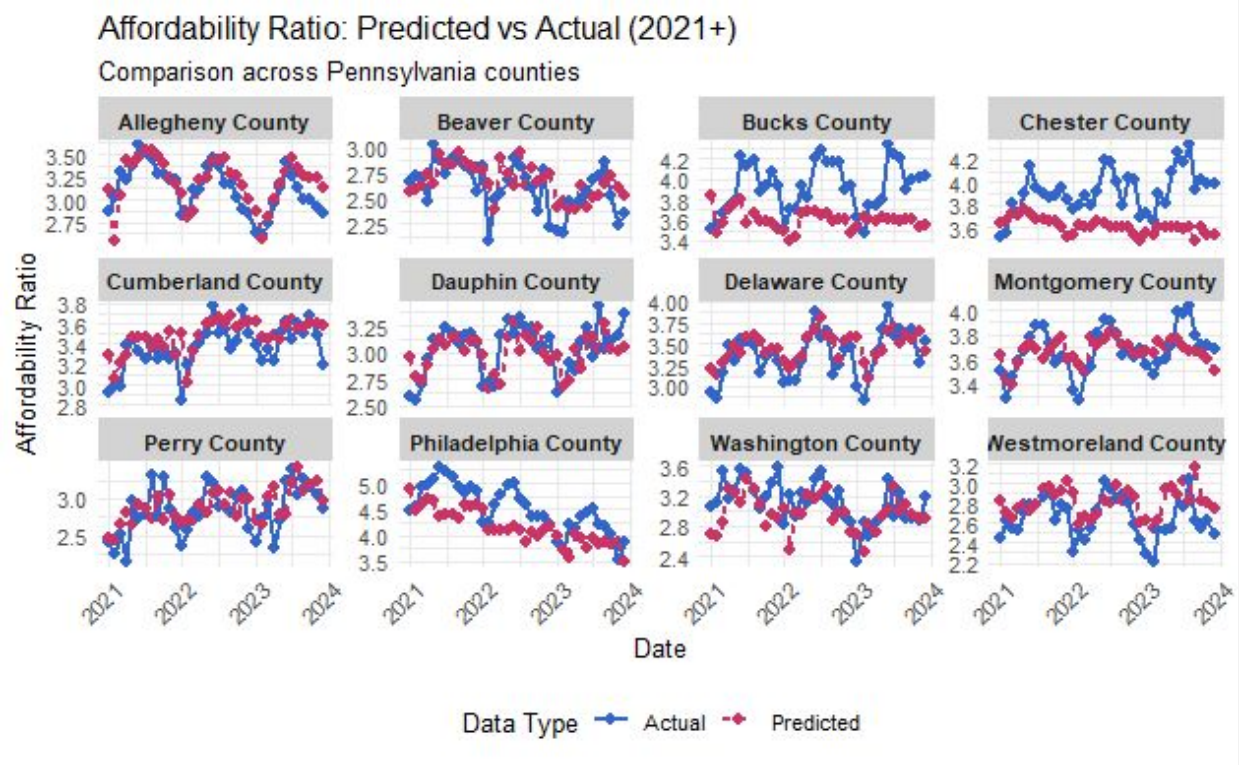
Part 5
Further Improvements and Results

Further Models: XGBoost

In addition to our ARIMA model we also tried out an XGBoost model for predicting affordability ratio. We will be leveraging it in the following cases:

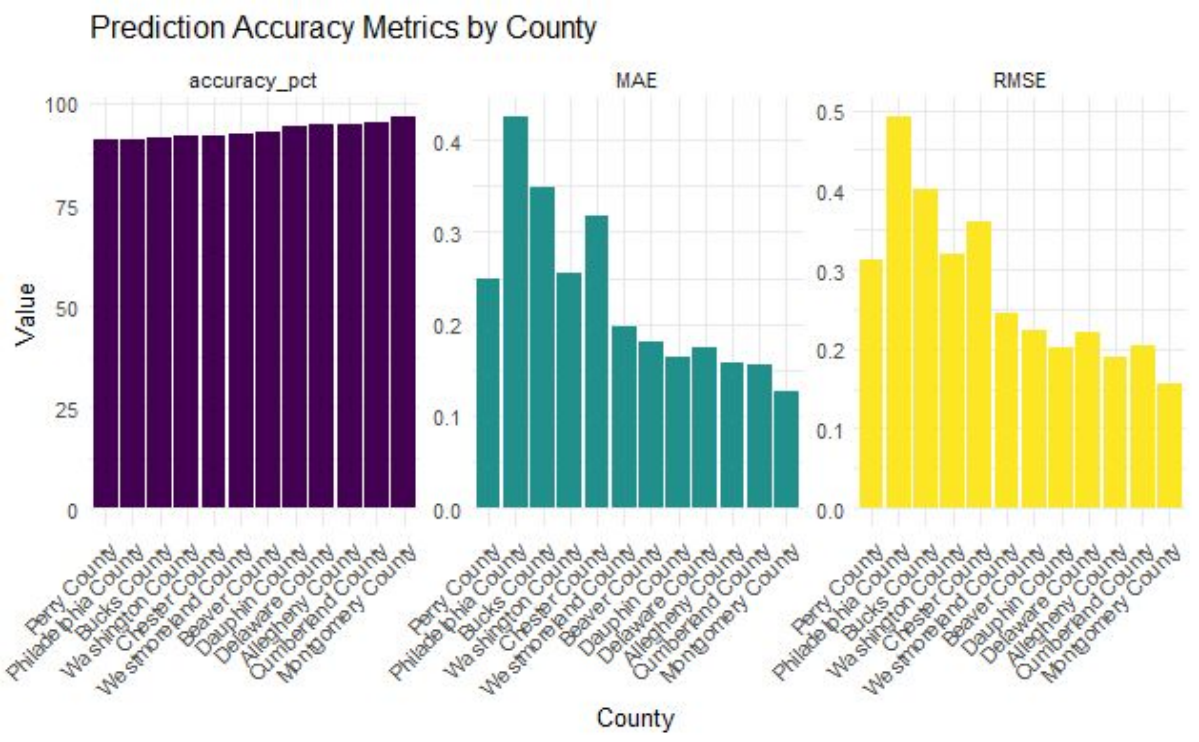
1. Use the model to predict the affordability for our previously mentioned metro areas of PA. (Pittsburgh, Harrisburg, and Philadelphia)
    a. It will use the current data from 2012 up to 2021 and then predict out to 2024
    b. This will allow us to learn the important predictors of the affordability ratio
    c. This will also act as our gauge for how well the model is doing so we can use the same settings to predict out data to 2028
2. Use the model to predict the affordability ratio of the counties out to 2028
    a. This will give us a idea of where the markets are going

XGBoost 2021-2024 Predictions



- The initial model we used was just a simple XGBoost model with lagged predictor for sale price of homes and incomes.
- It performed well but we can use more variables and tune our parameters to perform better

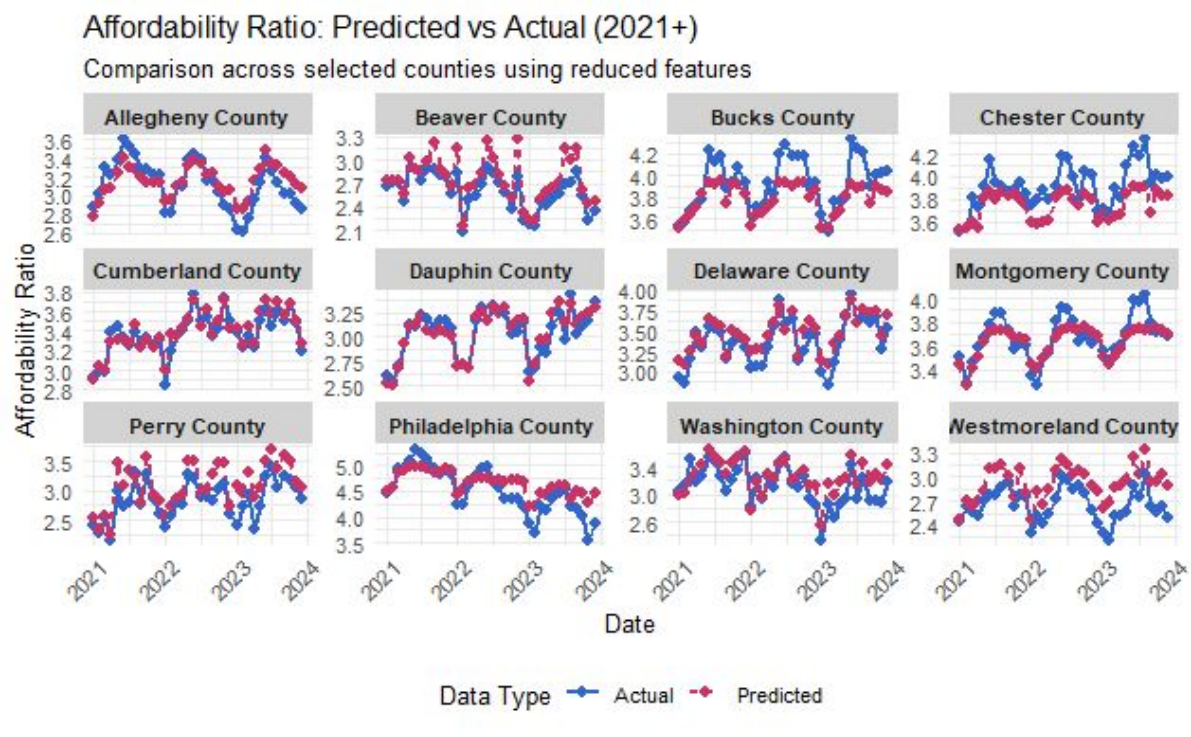
XGBoost Evaluation Metrics



How we will evaluate the model:

1. Accuracy: The proportion of correct predictions out of the total predictions made
2. MAE (Mean Absolute Error): measures the average magnitude of errors in predictions, without considering their direction
3. RMSE (Root Mean Squared Error): calculates the square root of the average squared differences between predicted and actual values

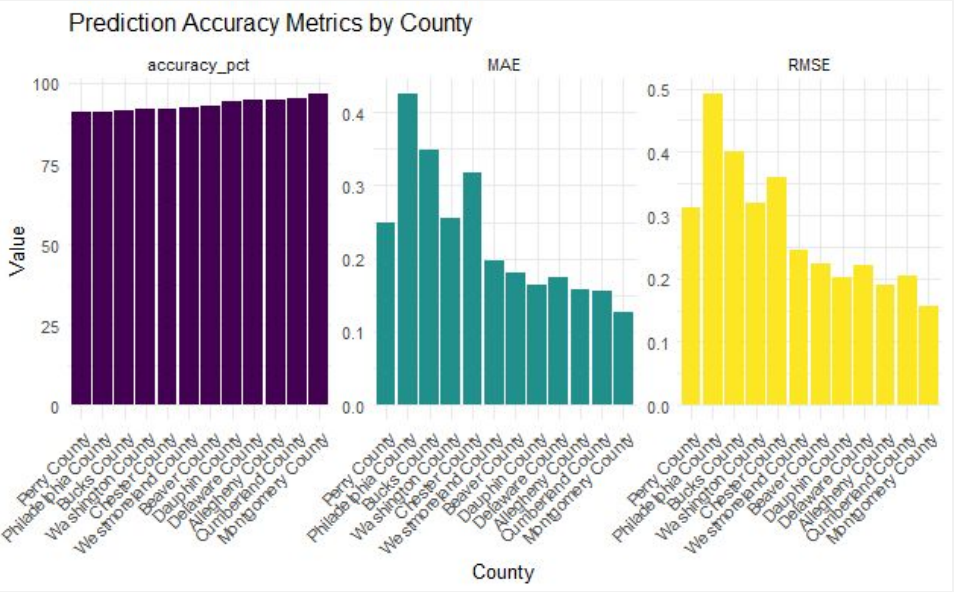
The Refined XGBoost Model



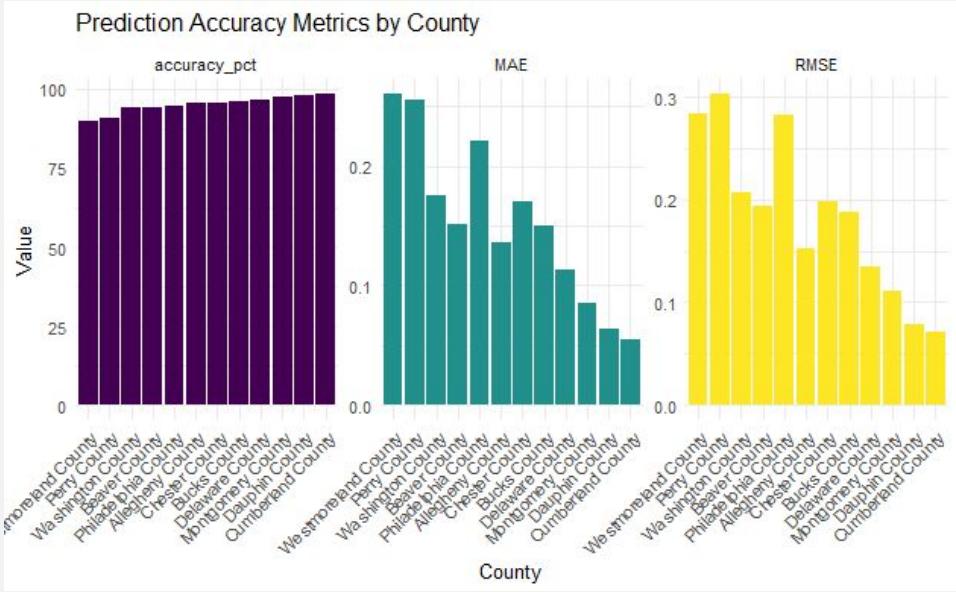
The final model that we derived used a variety of tuning methods to improve the results of the original XGBoost model:

1. Tuned hyper parameters of the model through a gridsearch
2. Introduced all features and then cut it down through the feature importance rankings
3. Adjusted the amount of time series data fed into the model
4. Leveraged lagged features

Refined XGBoost Results



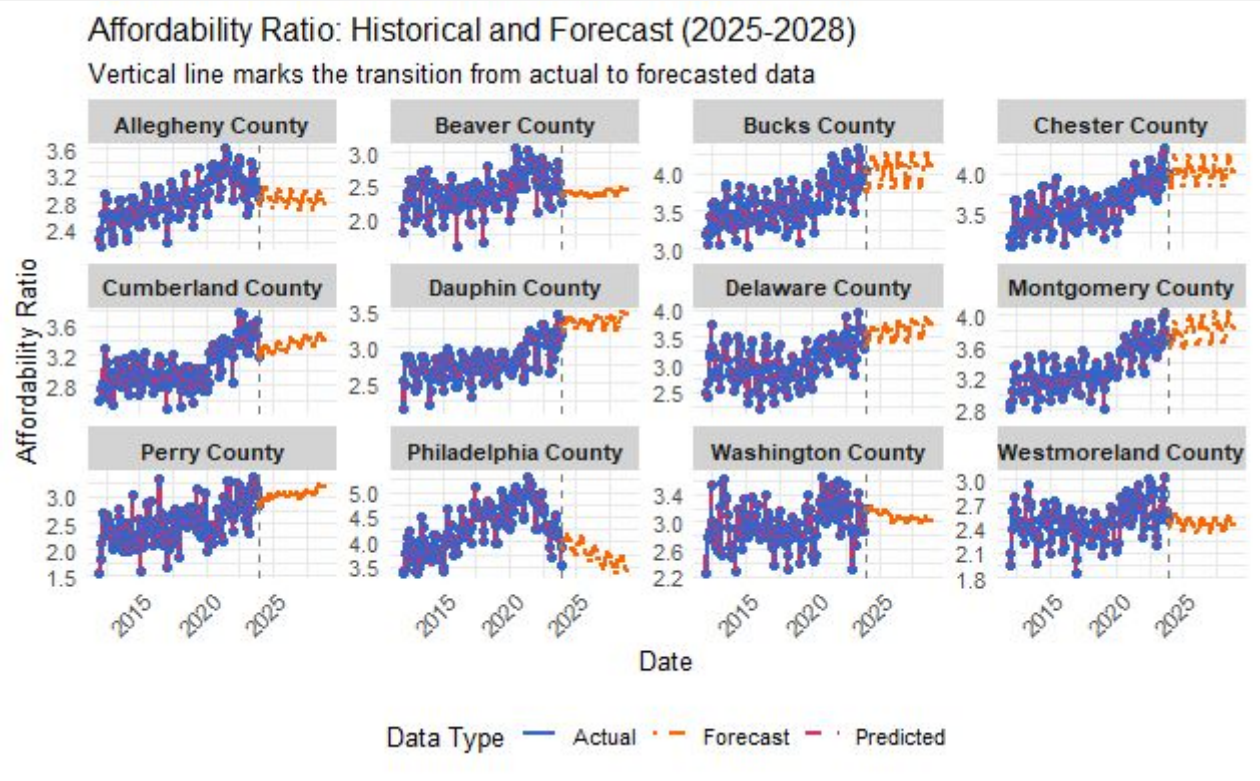
The Original Version



The Refined Version

- The Refined Version outperformed the original version in all areas and will act as our outline for a model to predict out to 2028

XGBoost 5 Year Predictions



- We then leveraged what we learned for refinement for the 2021-2024 predictions to make the model to predict out to 2028
- We fed all the data we have from 2012 to 2023 to get the most accurate results

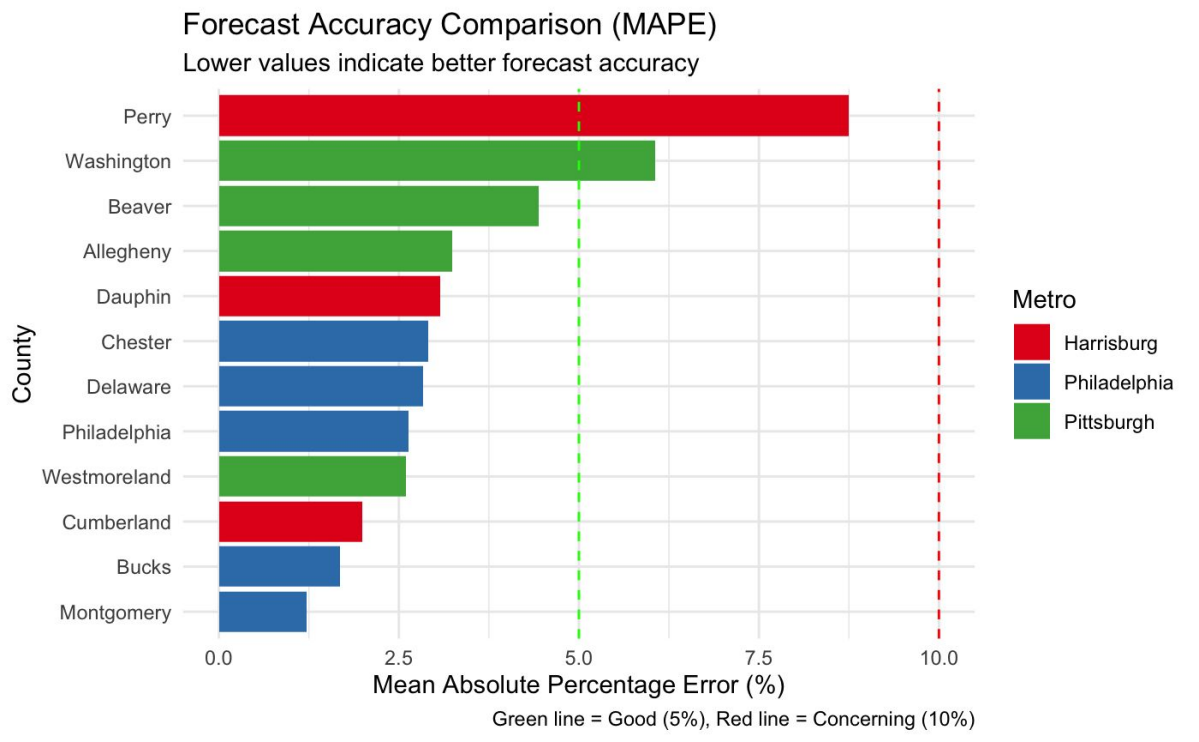
Our Plans for the Final Week

Our final steps to get to the final finished product are:

1. Finish refining the models
    a. ARIMA
    b. XGBoost
        i. We may make unique features for areas that have sub par predictions on the 2021-2024 predictions we made
2. Make Evaluation Metrics to compare the two
3. Determine the most reasonable estimate of the affordability ratio
4. Deliver our recommendations for the most affordable areas for young people looking to buy houses in Pittsburgh, Philadelphia and Harrisburg over the course of the next 5 years
    a. This will also include reasoning from our EDA on all of the regions from earlier
    b. This will allow us to give a comprehensive suggestion to home buyers looking to get a quality yet affordable home

Part 6
Further Improvements and Results

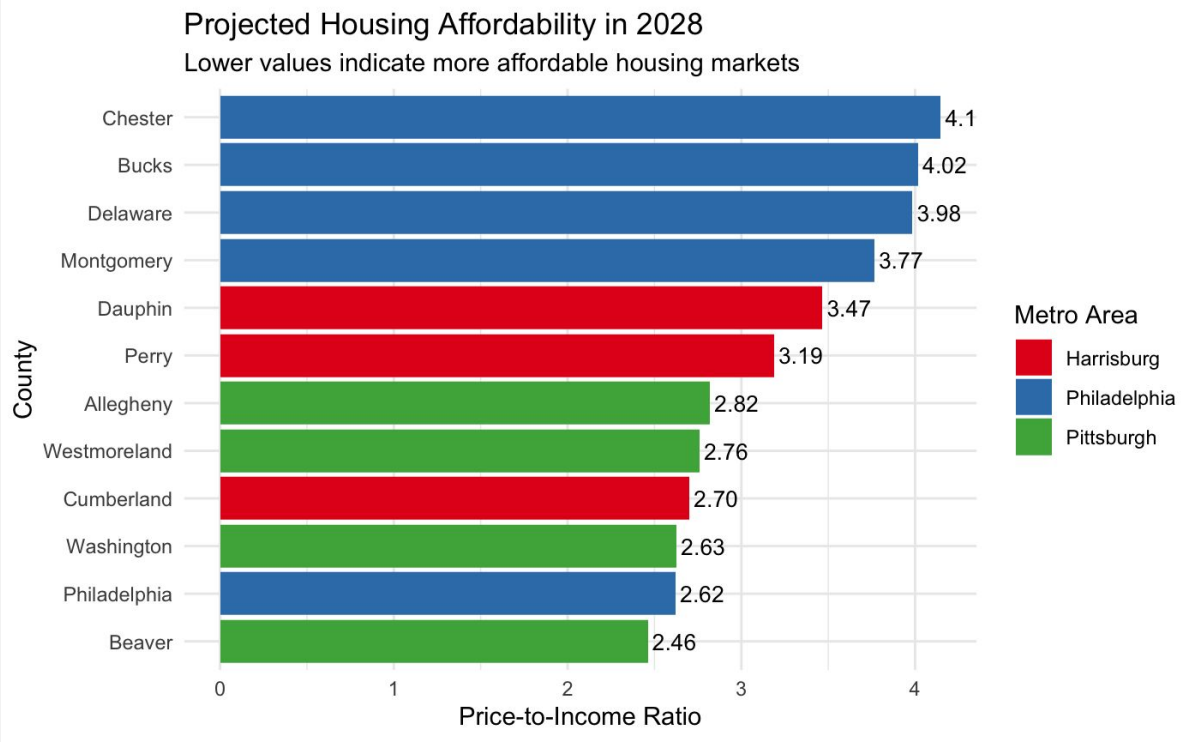
ARIMA Forecast MAPE



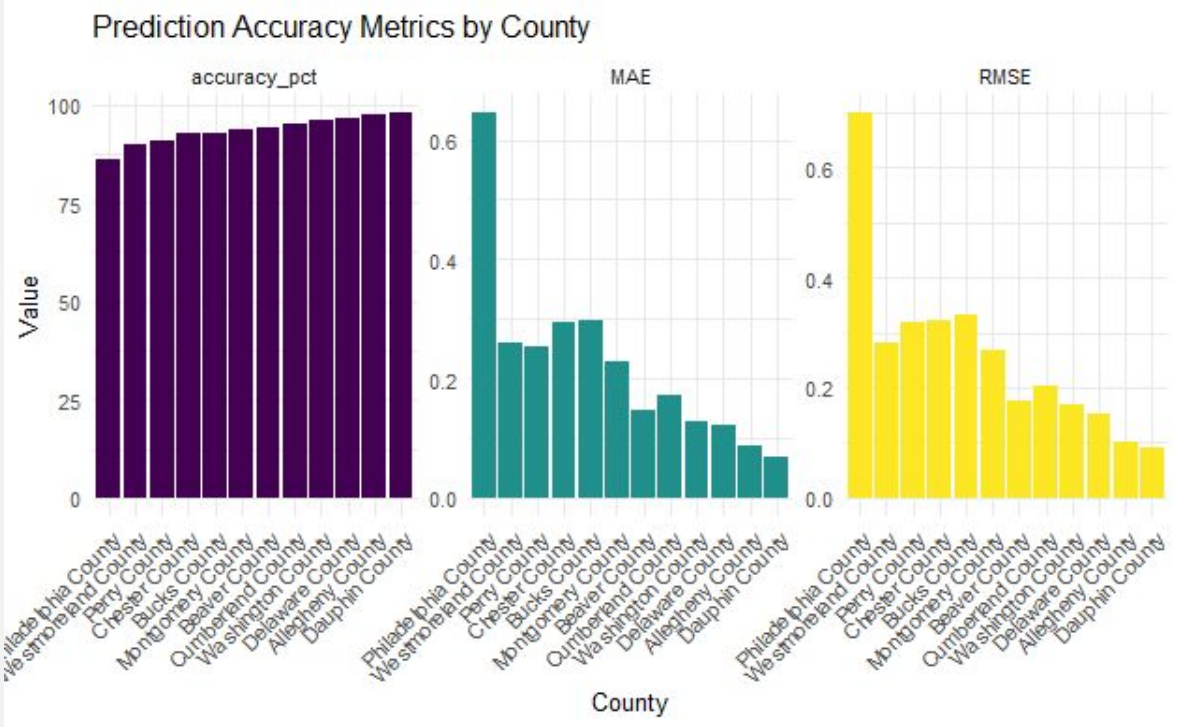
How forecasting improved?

- We forecasted each county 4 separate times using different arima specifications
  - Auto.arima, simple, simple w/ no differencing, handling drift
- Falls back to exponential smoothing if all ARIMA models fail (which none did)
- Returns the best forecast

Predicted Ratios (ARIMA)



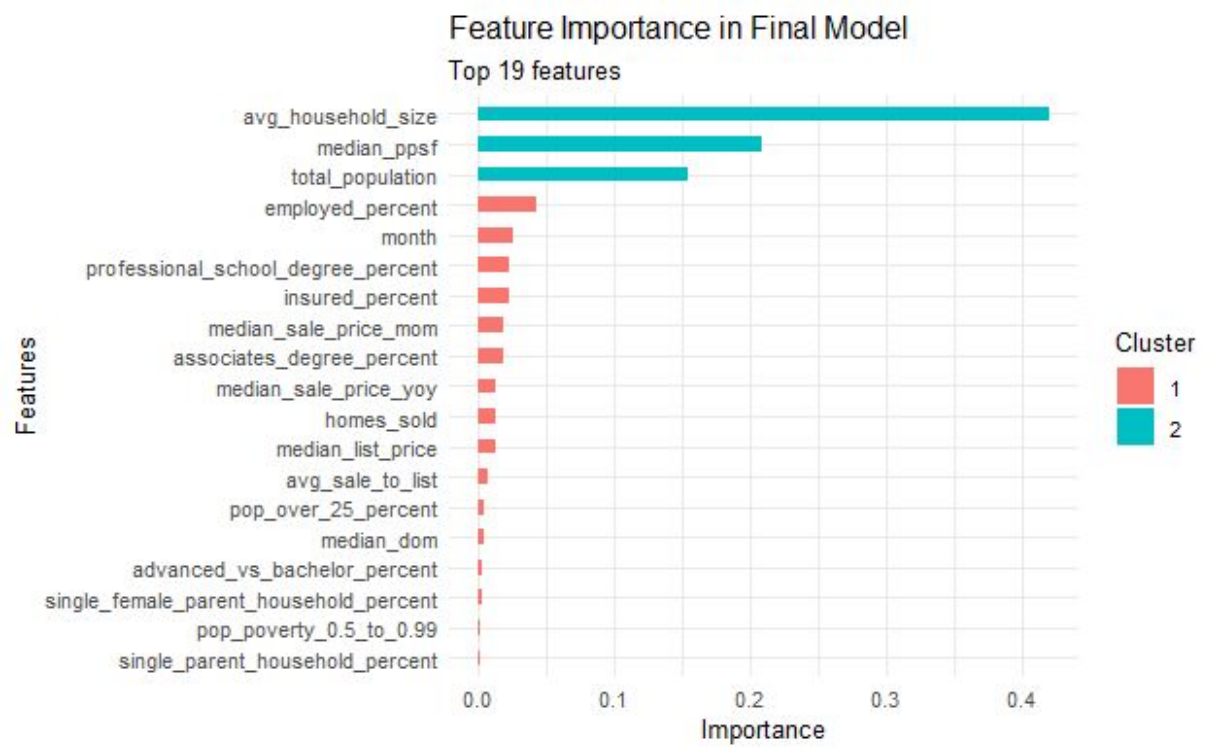
Boost Forecast Metrics



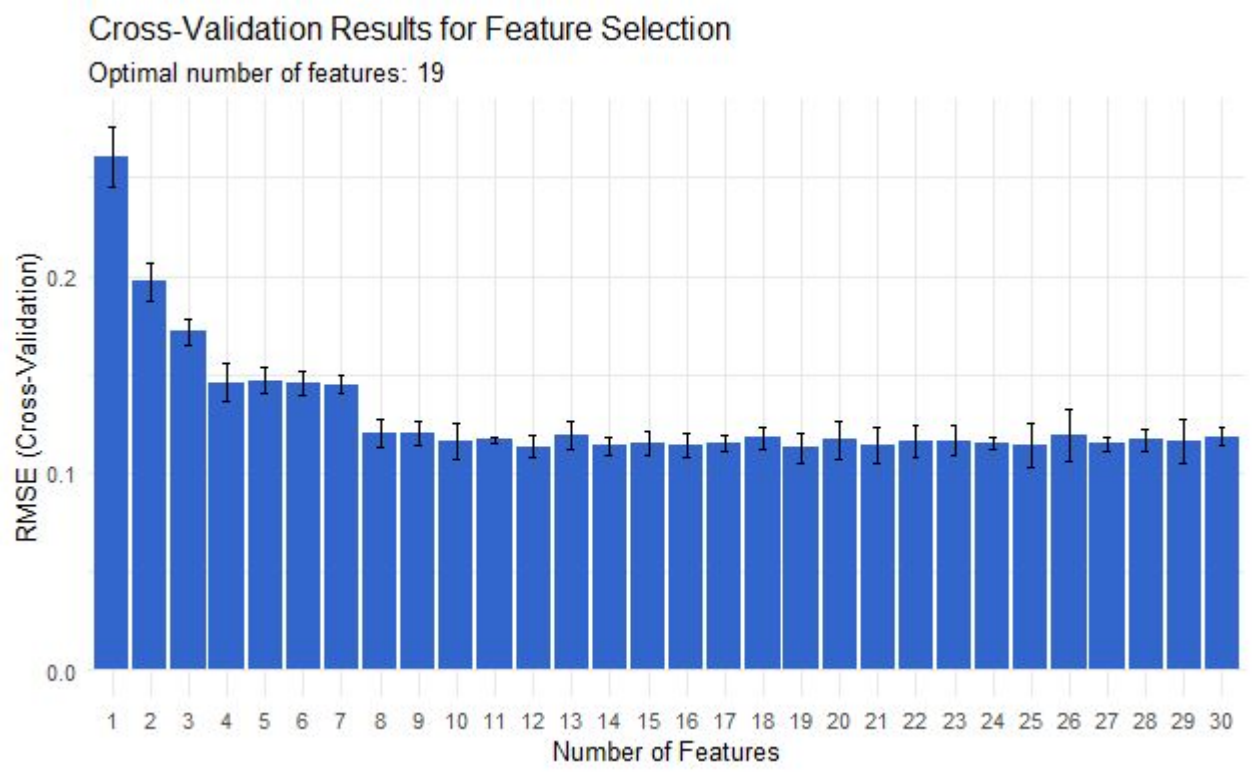
The Revised Metrics

- All metrics were able to improve besides those of Philadelphia County

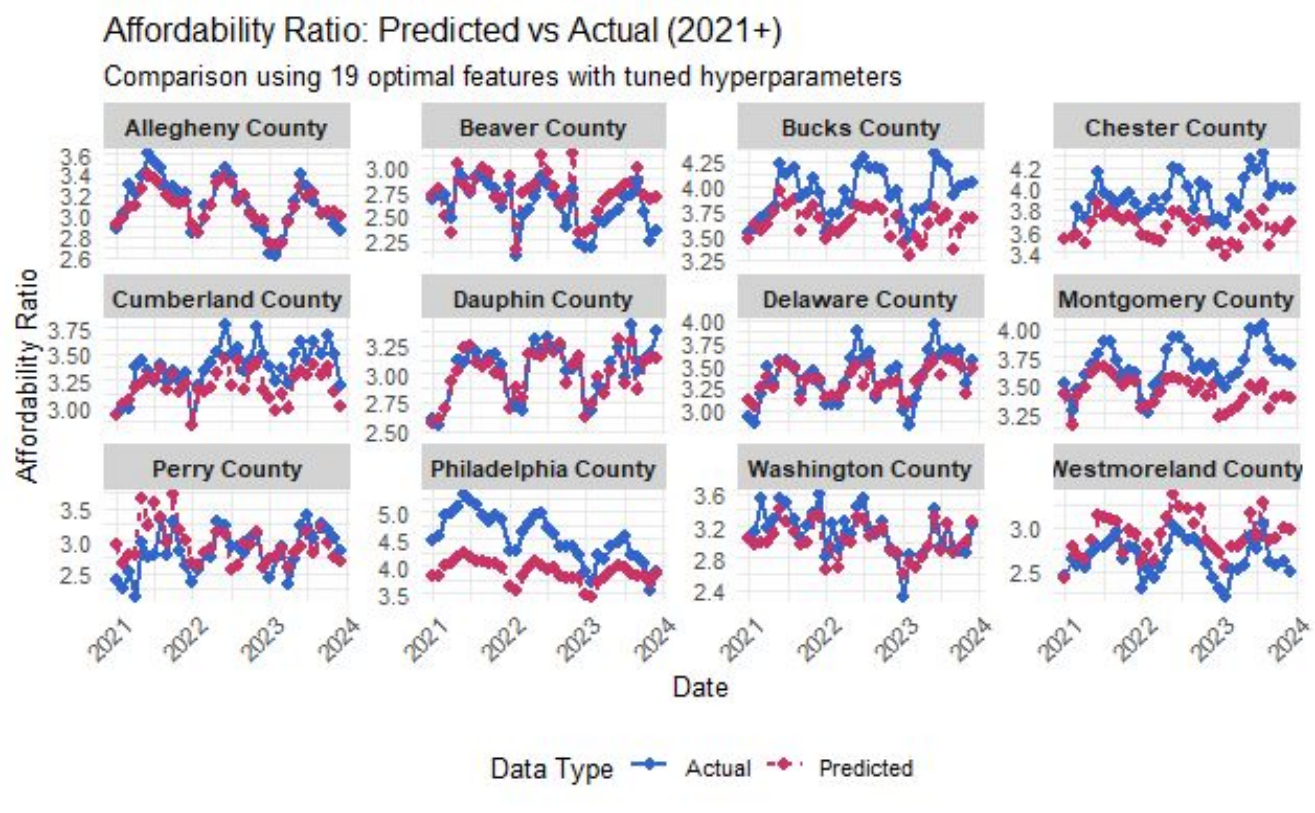
Top Features (Boost)



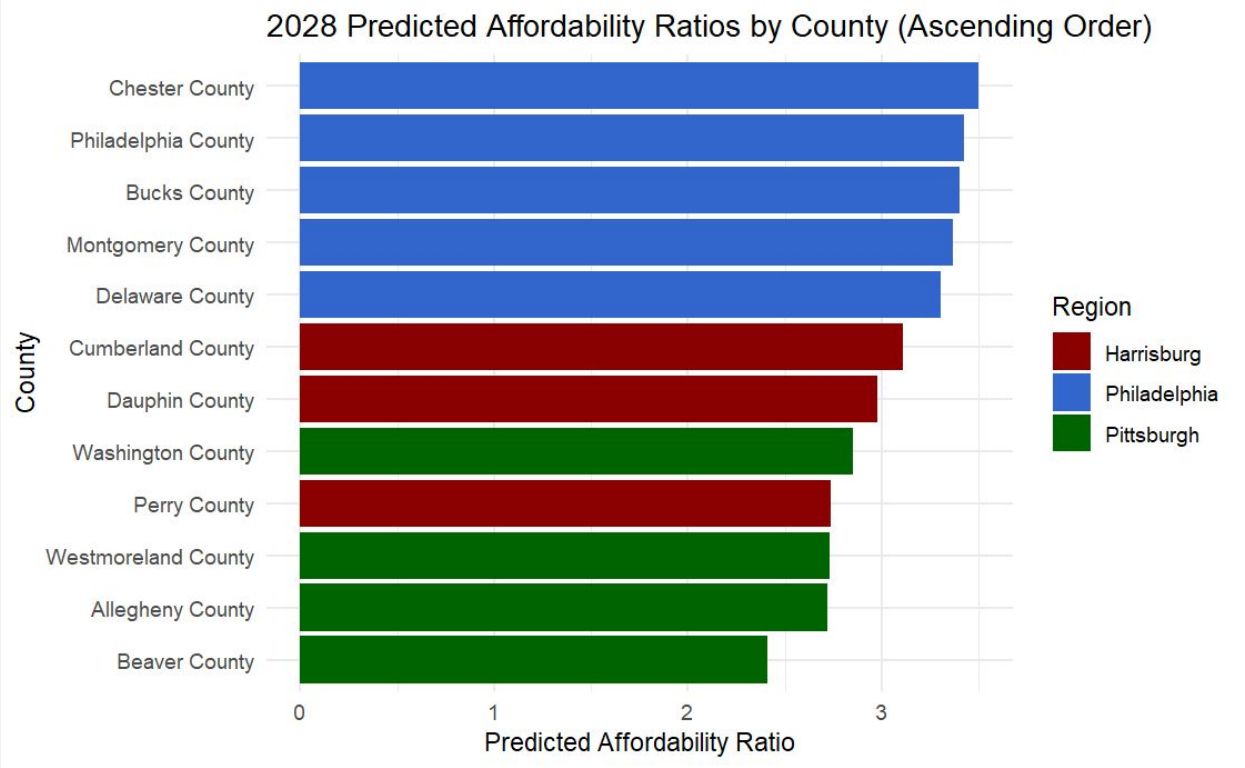
Top Features (Boost)



Boost Forecast



Predicted Ratios (Boost)



The Overall Rankings at 2028

- Overall the rankings are pretty similar to those of the ARIMA model
- The contradictions come with the troublesome county of Philadelphia
- The values are also slightly lower than the previous model