Aurelia Arnett

IST 718 – Big Data Analytics

**Introduction**

The pricing and value of the housing market can vary with time and greatly depend on the geographic region the house is based in. In some locations, the housing value fluctuates with time while in other locations the housing value remains stable and more predictable. The goal of this case study is to evaluate housing data for zip codes across the US, based on historical housing values, and recommend 3 zip codes for the Syracuse Real Estate Investment Trust to invest in. Through a series of modeling historical data and forecasting the housing value in the future, recommendations will be made based on both the predicted value change with time, whether that increases or decreases, and the potential error of the prediction, based on the range of the prediction.

**Initial Analysis**

The dataset provided includes the mean housing value for all homes per zip code from 1996 through March of 2020. To understand how housing value may change with time, the program computes the mean housing value of zip codes within four different metro areas in Arkansas (Hot Springs, Little Rock, Fayetteville, and Searcy) are plotted as a time series.

This method consisted of creating subset data frames from the zip code data, containing the zip codes in the four Arkansas metro areas, and then calculating the mean value of all zip codes of each metro area. One final Arkansas dataset was constructed, and the time series was printed:

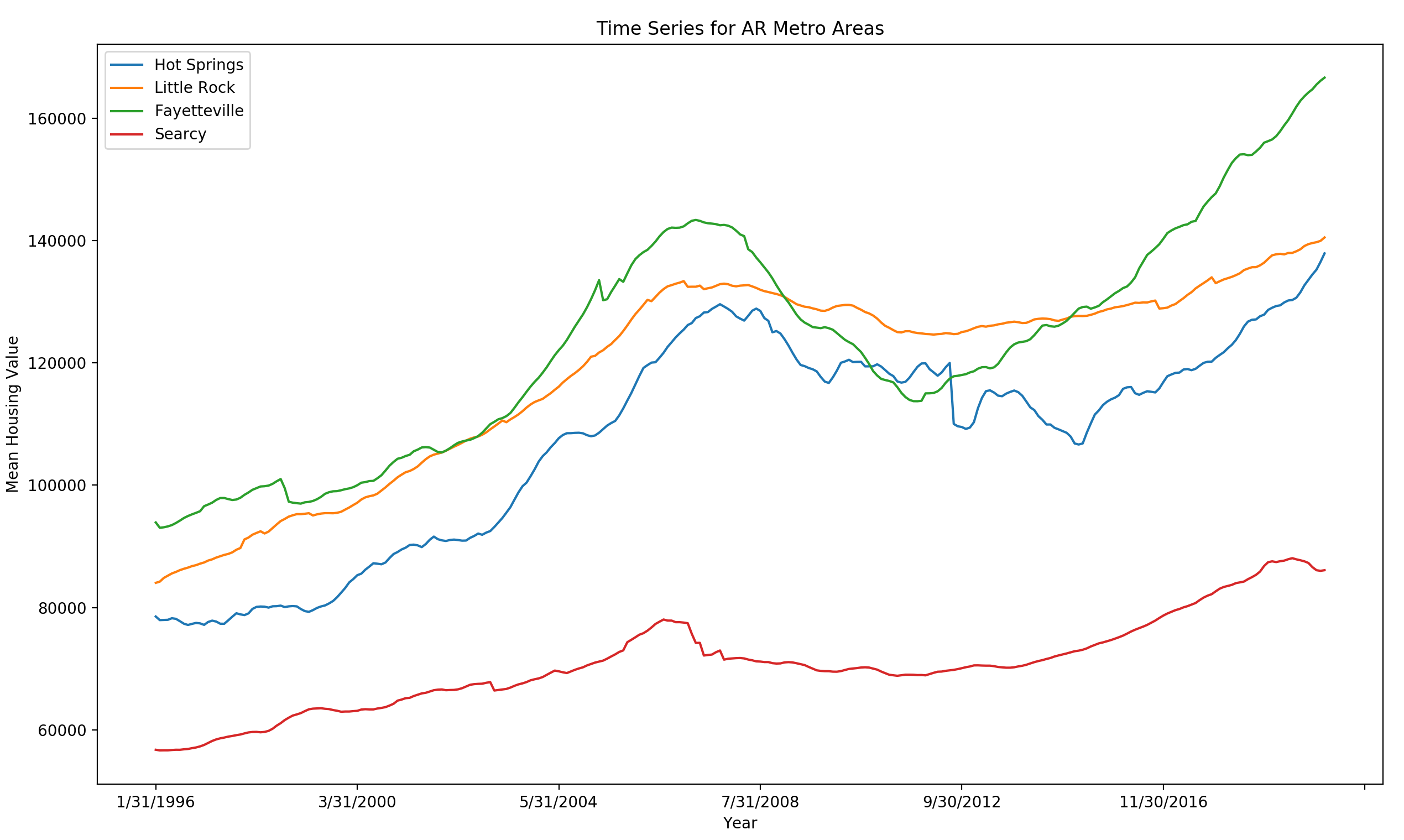


Figure 1: Time series for 4 metro areas in Arkansas

The time series represents the variability of housing value with time for the four metro areas. By focusing on the original and final housing value of all four areas, it can be concluded that the overall housing value for each metro area has increased over 24 years. In taking a closer look at the value for houses throughout time, we see this was not always the case especially for Hot Springs and Fayetteville as the value increases exponentially from Jan 1996 to about July 2008, and then dramatically decreases. Housing values in the Hot Springs area continue to vary until about 2014 and then we see the value dramatically increase again. These trends indicate that while the housing value shows a growth in March of 2020, it would be challenging to predict whether this growth for Hot Springs and Fayetteville is sustainable with time or if the value would drop. Houses in Fayetteville do show a little more predictability after about 2010, with a value increase, however, show a high entry point for those considering buying in Fayetteville today. Housing values in Searcy show very little variability with time, indicating low risk and high predictability, however shows very little growth and potentially indicating very little interest exists to buy in that area.. This may be a great investment for those who are looking for stability in housing value, though must consider other factors such as environment, walking score, ease of access to entertainment and resources, and things to do in the area.

To further understand the how the housing value may vary with time, the program next builds prediction model applying the Prophet method on these zip codes from the previous 5 years, forecasts the housing value 1 year forward, calculates the error of the model by subtracting the lower prediction value from the upper prediction values, and stores this error term. The prediction error is then plotted and returned:

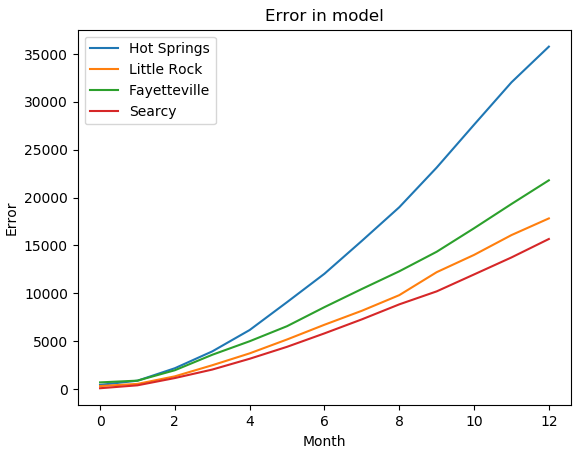


Figure 2: Error prediction model of Arkansas zip codes over next 1 year

The prediction error plot confirms the initial analysis that the housing value in Searcy is the most predictable of housing values in these Arkansas areas and can be predicted with low risk. Hot Springs on the other hand shows a dramatic increase in prediction error, especially the farther out you are trying to forecast, which indicates high risk and variability for investment over time. The value of housing in Fayetteville and Little Rock show higher error and risk than Searcy, but doesn’t have as dramatic of an increase in prediction error over time as Hot Springs.

Next, the program will evaluate zip codes in the greater Seattle area, specifically looking at King County, as location is important to me and my first choice of investment would be to look into houses in my home area.

**Further Analysis – King County, WA**

For the King County analysis, risk and forecast terms are compared relatively, meaning mid-level risk means mid-level compared to the error term of other zip codes.

To understand the housing market in King County, the program creates a dataset containing zip codes only in the King County area and then applies the Prophet model using 10 years of historical housing values to the dataset. Two forecast models are created and plotted to show the potential growth and risk over the next 1 year and 5 years. The program again stores the error term and prints out the zip codes that rank with the top 5 lowest errors with time. The first two plots printed evaluate value growth over the next 1 year:

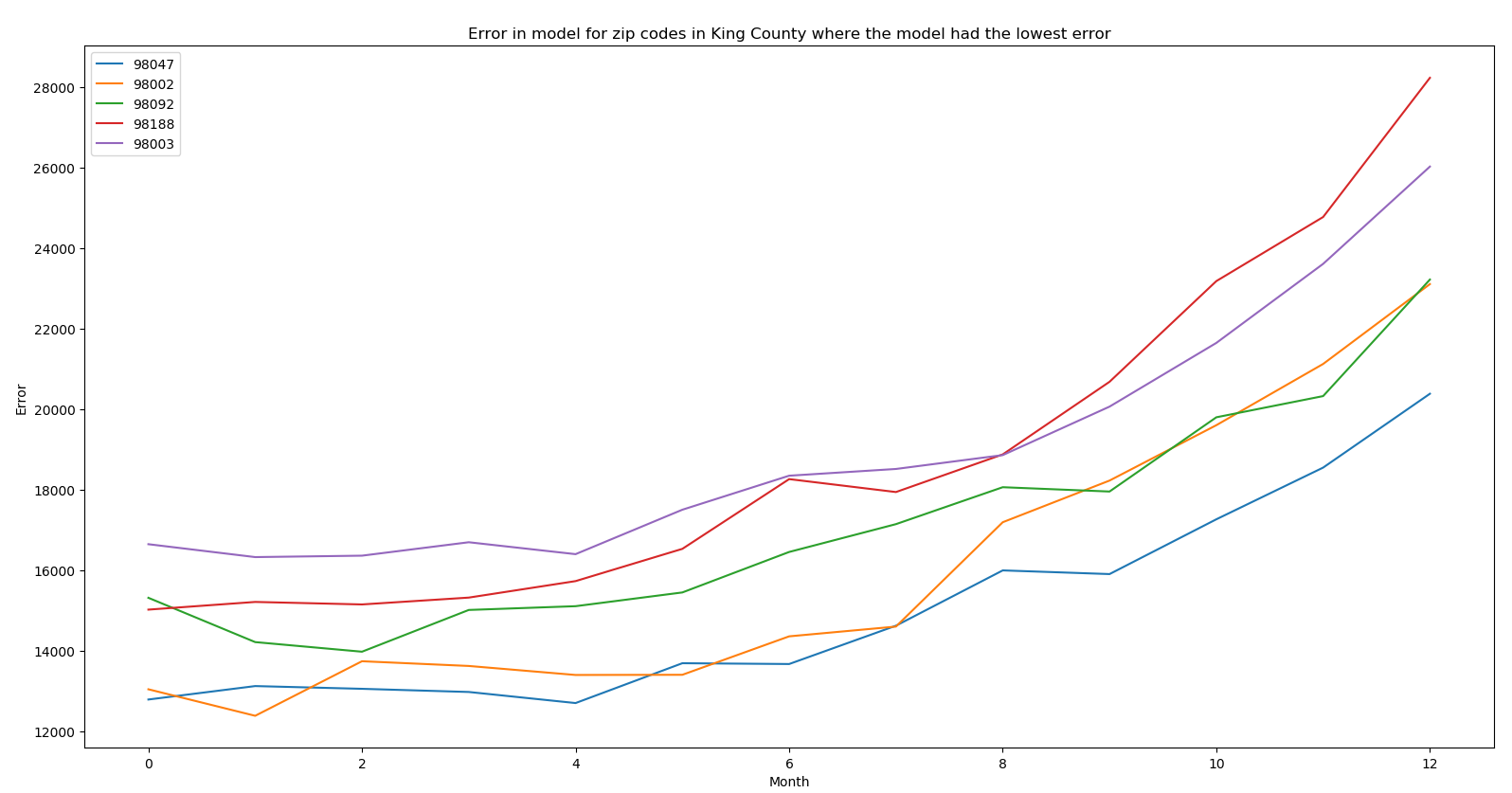


Figure 3: Error prediction model of zip codes with the lowest error, forecasted over next 1 year

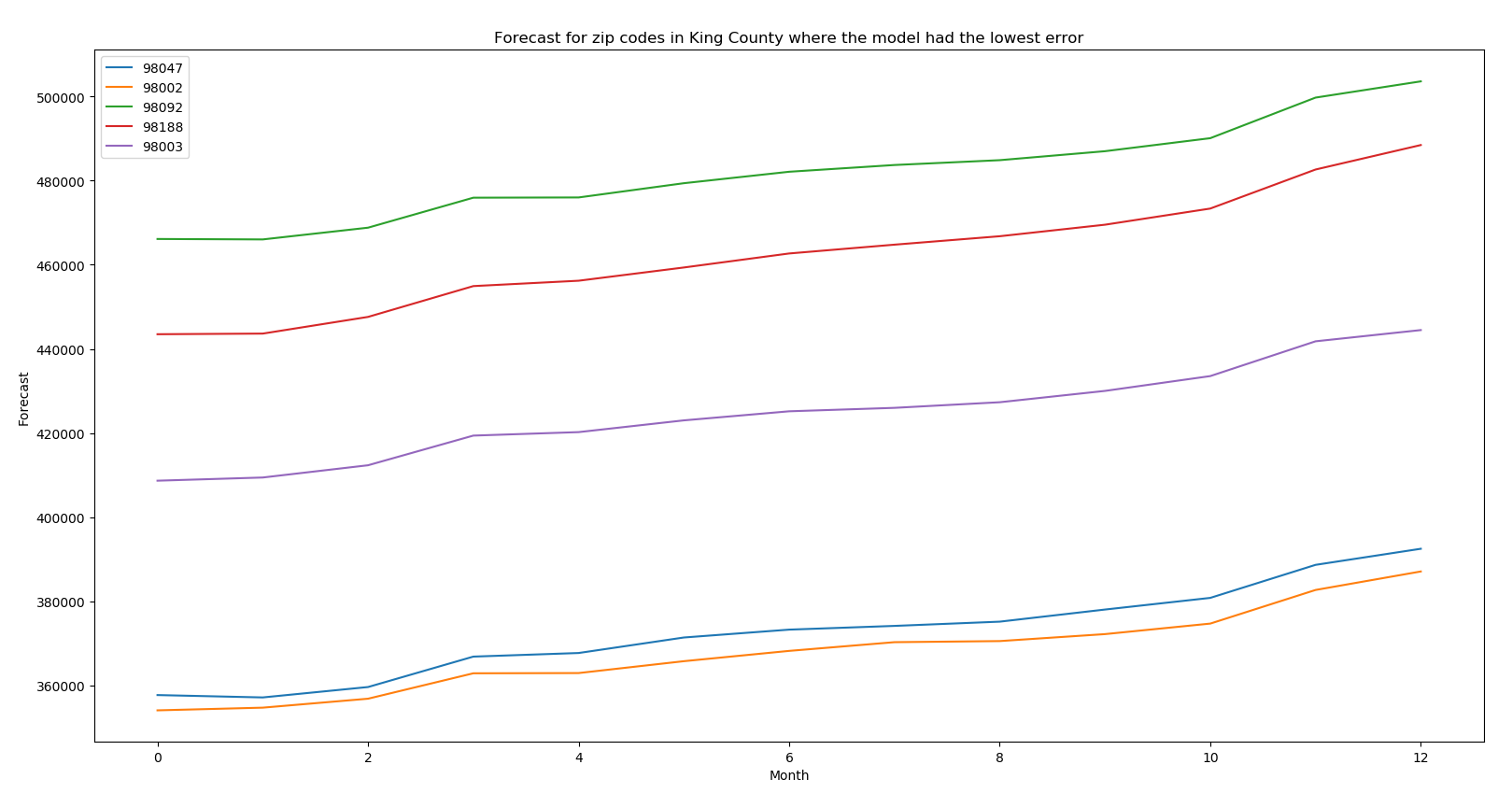


Figure 4: Predicted growth model of zip codes with the lowest error, forecasted over next 1 year

When the model forecasts the housing value over the next 1 year, there is generally the lowest error with time and some housing growth for zip code 98047, which would be the best investment in the King County error without considering other factors. Houses in this zip code also show to have a low-price entry point, making this the most cost-effective area of investment. On the opposite spectrum, zip codes 98188 and 98003 show to have higher growth with a lot of variation in error over 1 year. Zip code 98092 seems to have mid-level risk, showing a subtle decrease in error term in the first few months before the error term begins to grow with time, and higher housing growth with time. A forecast on zip codes in King County over the next 5 years is plotted to evaluate whether these results are consistent with time.

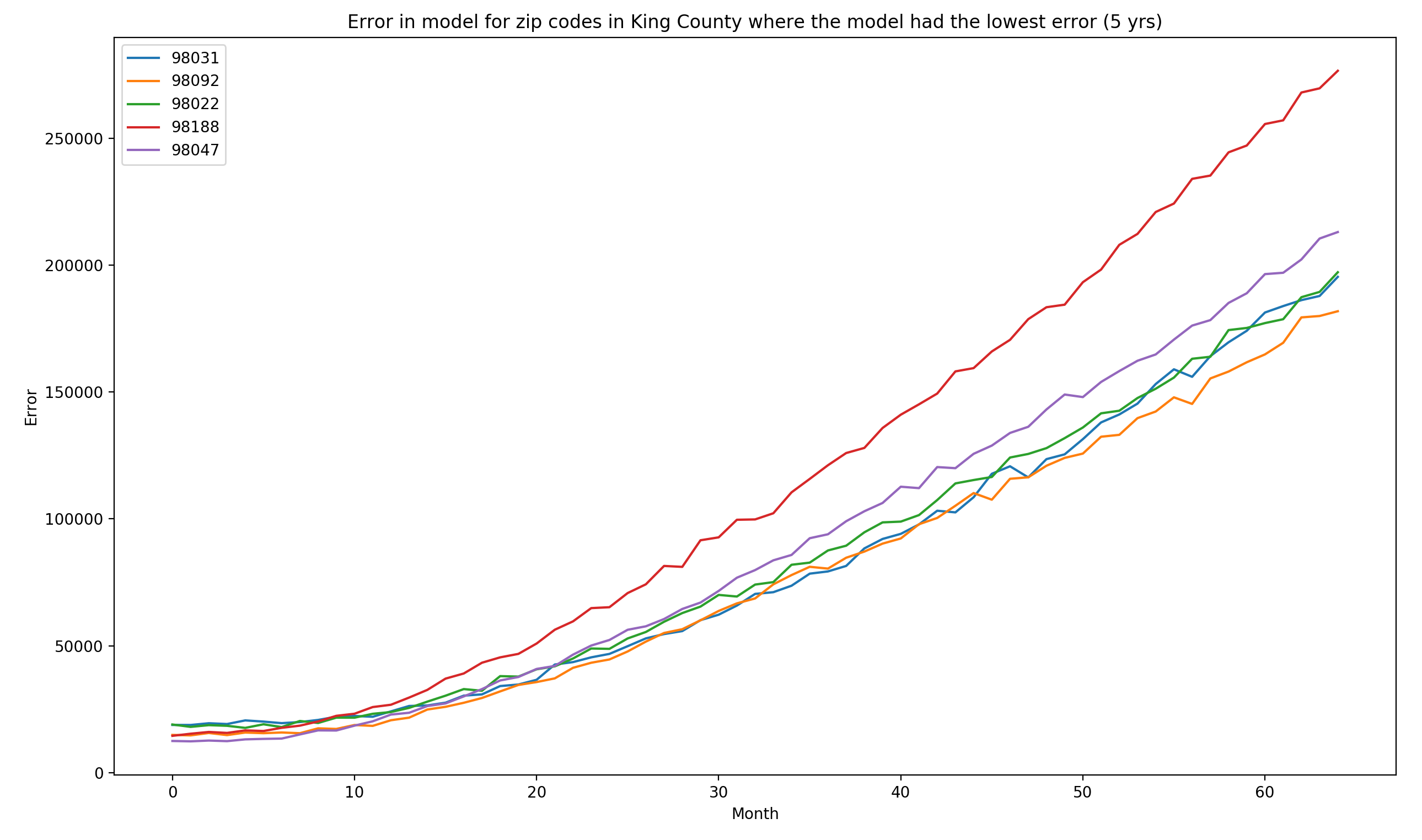


Figure 5: Error prediction model of zip codes with the lowest error, forecasted over next 5 years

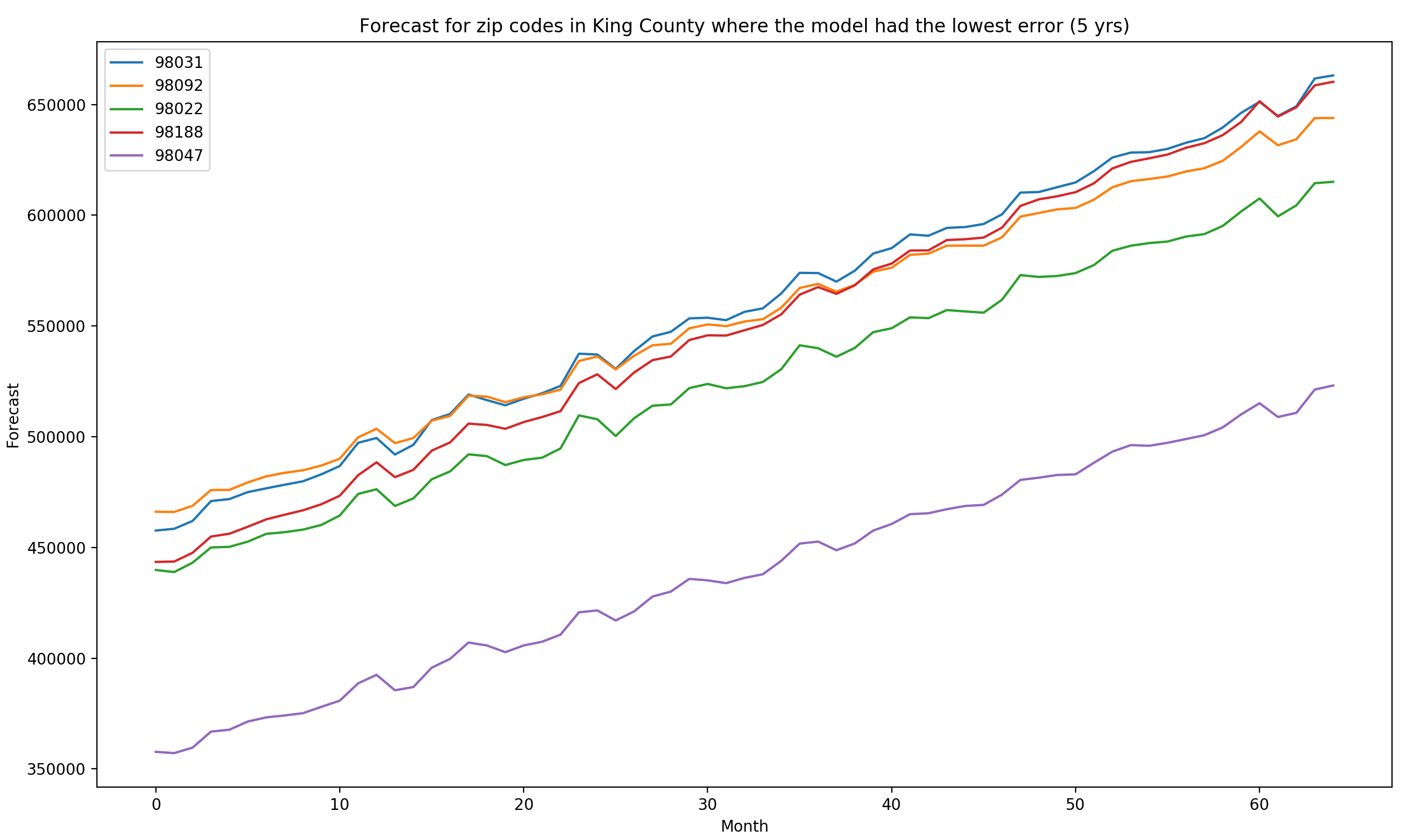


Figure 6: Predicted growth model of zip codes with the lowest error, forecasted over next 5 years

When the model forecasts the housing value over the next 5 years, we see only 3 of the original 5 zip codes with the lowest error over 1 year is returned: 98047, 98118, 98092. This indicates that while risk in predicting the housing value itself may vary with time but the ability to model the housing value with time becomes more predictable than other zip codes in King County.

In closely evaluating each plot for error in model and housing forecast, zip code 98118 continues to show both high risk and high growth with time, zip code 98047 shows to have a little bit higher error though a lot more growth, and zip code 98092 show to have the lowest risk and high growth with time. The other two zip codes, 98031 and 98022 seem to have the same mid-level of risk with exponential growth over time.

When it comes to investing in the King County area, I would choose to invest in zip code 98092 because it was shown to have one of the lowest error terms based on the two models analyzed in this case study, shown to have lower error with time in the 5 year forecast, and shown to have fairly stable housing value growth over 1 year and over 5 years.

**Analysis on all zips**

The final component of this case study analyzes the dataset consisting of all zip codes to recommend 3 investment areas for the Syracuse Real Estate Investment Trust. The program looks at historical housing data from 10 years prior, starting with March 2010, and forecasts the value over the next 10 years. The process begins with dropping all zip codes in the dataset that didn’t have a housing value starting in 2010, as the Prophet model won’t run on zip codes that have multiple instances of missing data, which reduces the dataset from about 30,000 zip codes to about 22,000. The program next divides the dataset into 4 datasets of 5,000 zip codes and 1 dataset of 2,000 zip codes and begins to run the Prophet model on each individual dataset. For each run, the program stores the error associated with forecasted values and the future predictions, adds the zip code label by indexing by the original train dataset, and exports the saved values into individual csv files. Once the program completes this process for each of the 5 datasets, all new csv files are imported into the program and appended into two new datasets: the first containing the approximate error associated with zip code forecasts over the next 10 years and the other containing the forecasted housing value for each zip code over the next 10 years. The final outputs are printed and the mean values for each zip code is calculated:

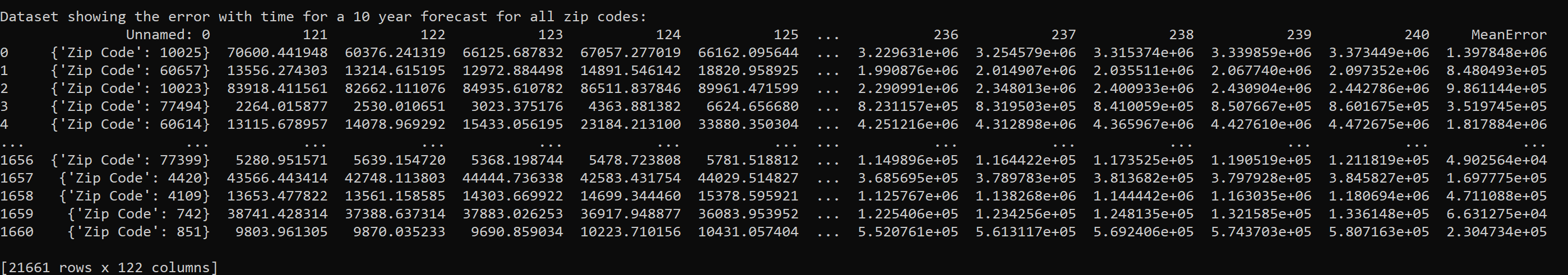


Figure 7: Final dataset revealing the predicted error for all zip codes over the next 10 years

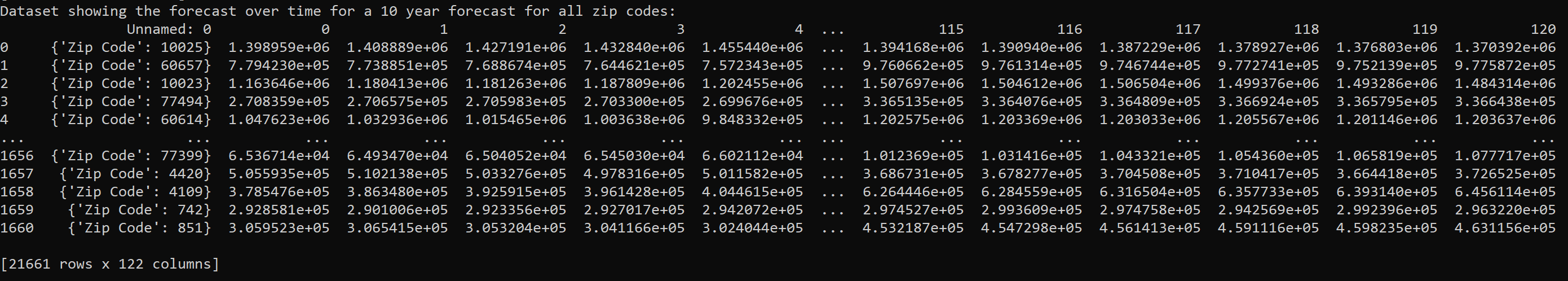


Figure 8: Final dataset revealing the forecasted housing values for all zip codes over the next 10 years

Finally, the program sorts the dataset containing zip code error by zip codes that are predicted to have the lowest error, indicating higher predictability and lower risk, and prints out the top 5 zip codes with the lowest error: 17976, 24726, 73438, 42032, and 24606:

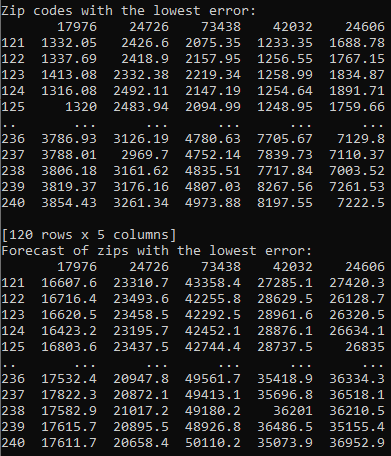


Figure 9: Top five zip codes predicted to have the lowest forecast error for housing value

The program then plots the error term and forecasted housing values for these zip codes over the next 10 years:

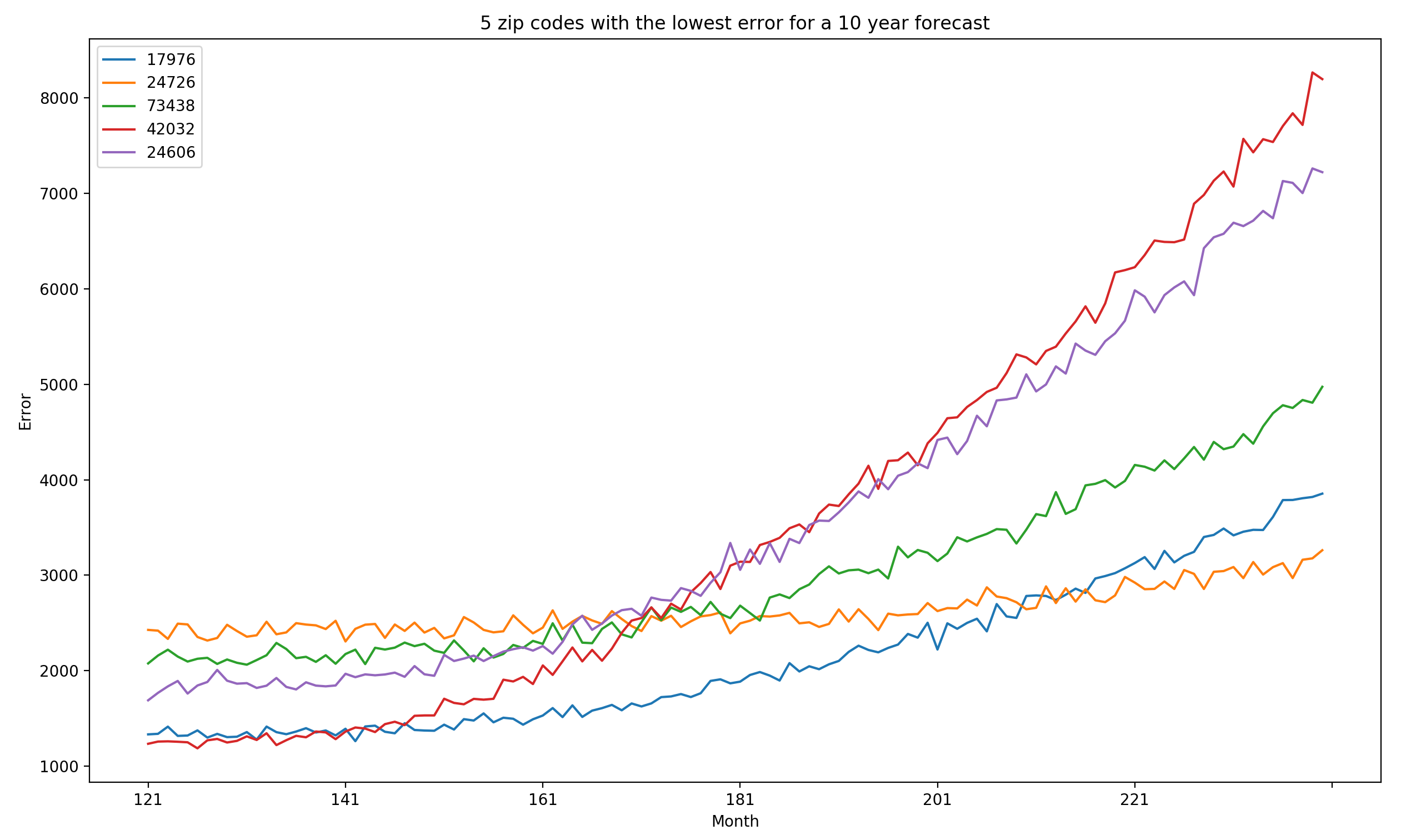


Figure 10: Prediction error for zip codes with the lowest error over the next 10 years

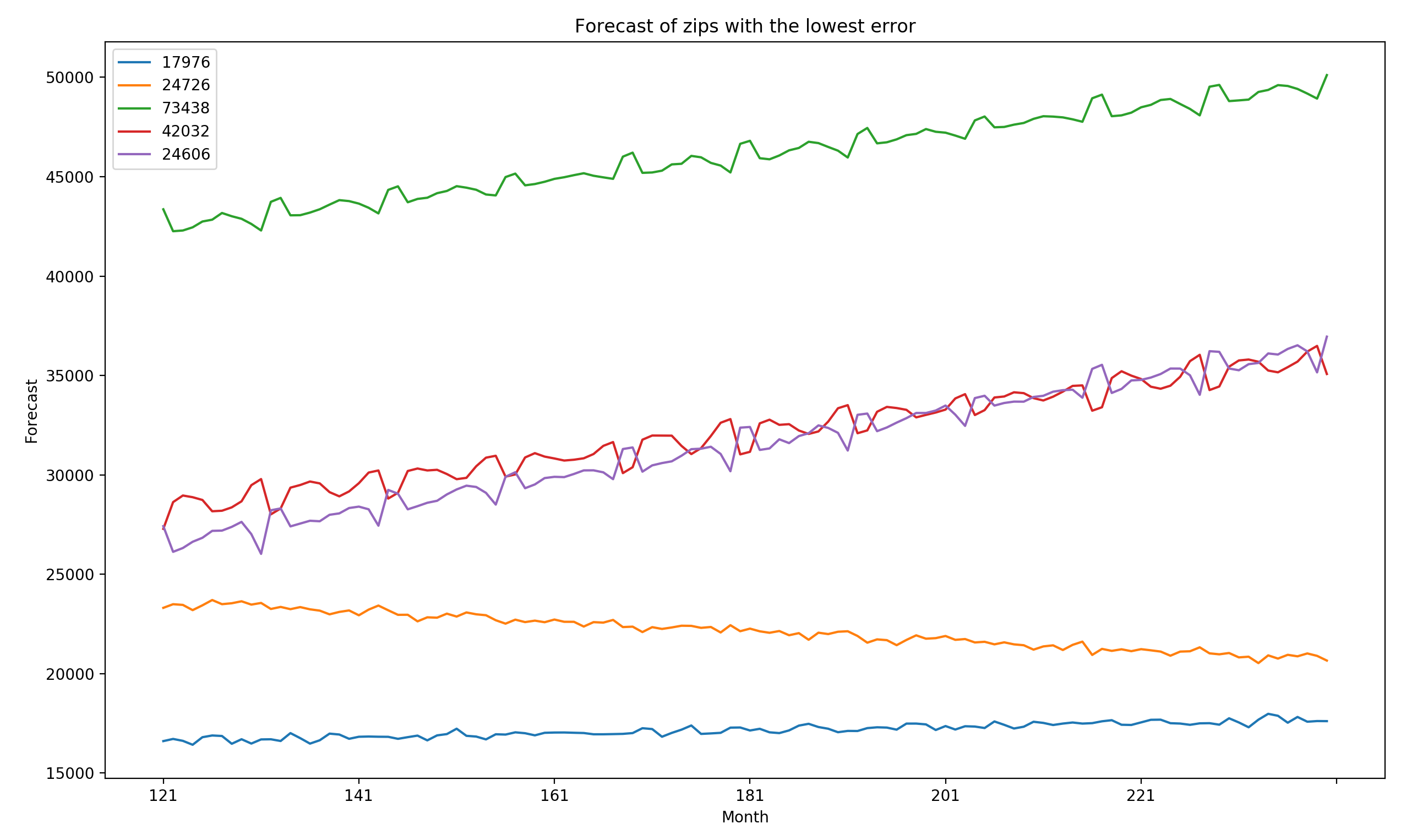


Figure 11: forecasted housing value for zip codes with the lowest error over the next 10 years

**Recommendation**

The Syracuse Real Estate Investment Trust should invest in zip code 73438, 42032, and 24606. When price and predicted growth is the highest consideration, 73438 provides the best value of all zip codes in the US because it shows to have overall growth and mid-level to low error among all zip codes in the dataset. While the other two zip codes show to have higher error relative to these 5 zip codes, they still have the lowest error among all zip codes within the US and consistent housing value growth over the next 10 years. The other two zip codes in the top 5 do show very little error however show no to negative growth over the next 10 years and would not be recommended areas of investment.

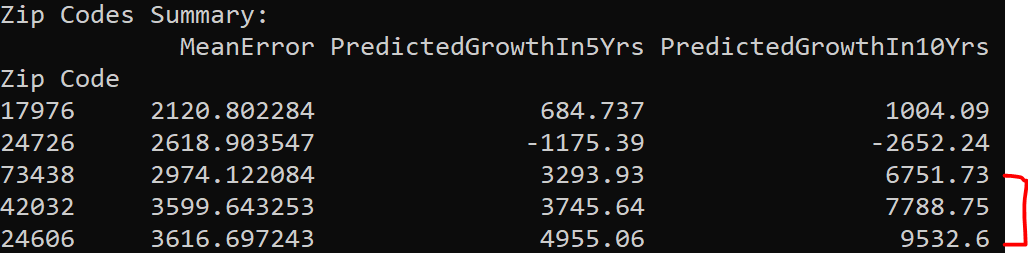
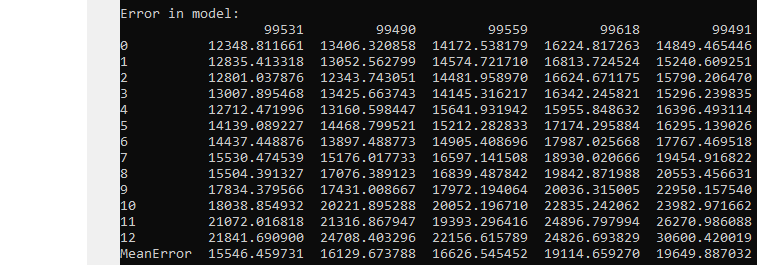


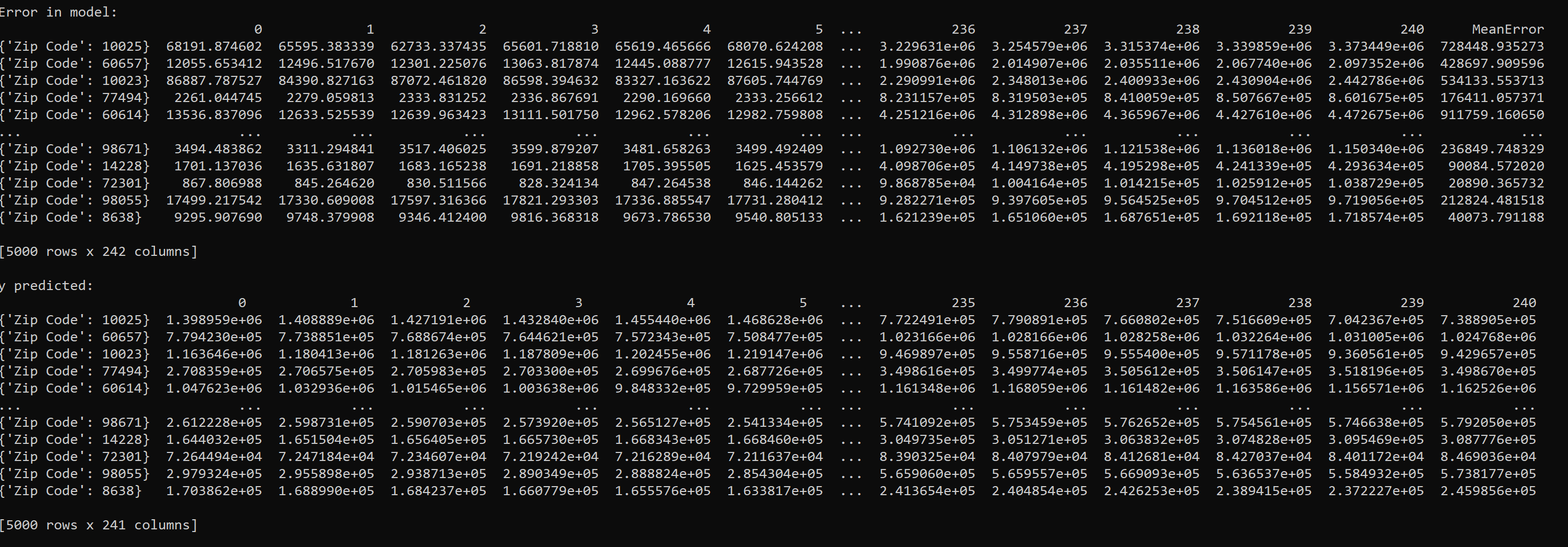
Figure 12: Zip code summary for the top 5 zip codes with the lowest error

**Appendix**: Graphics

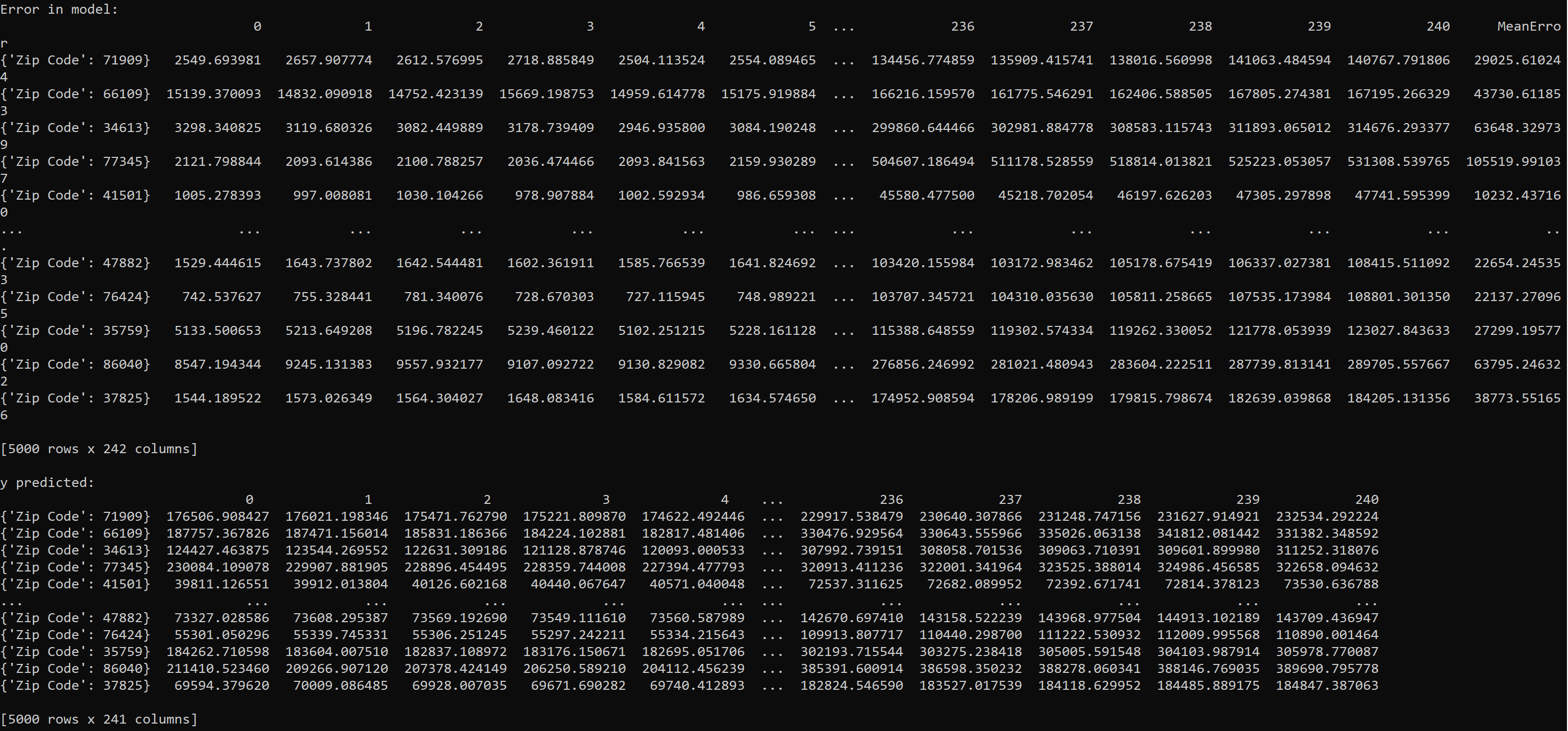
|  |  |
| --- | --- |
| Forecast model of zip codes in Hot Springs, AR | Forecast model of zip codes in Little Rock, AR |
| Forecast model of zip codes in Fayetteville, AR | Forecast model of zip codes in Searcy, AR |



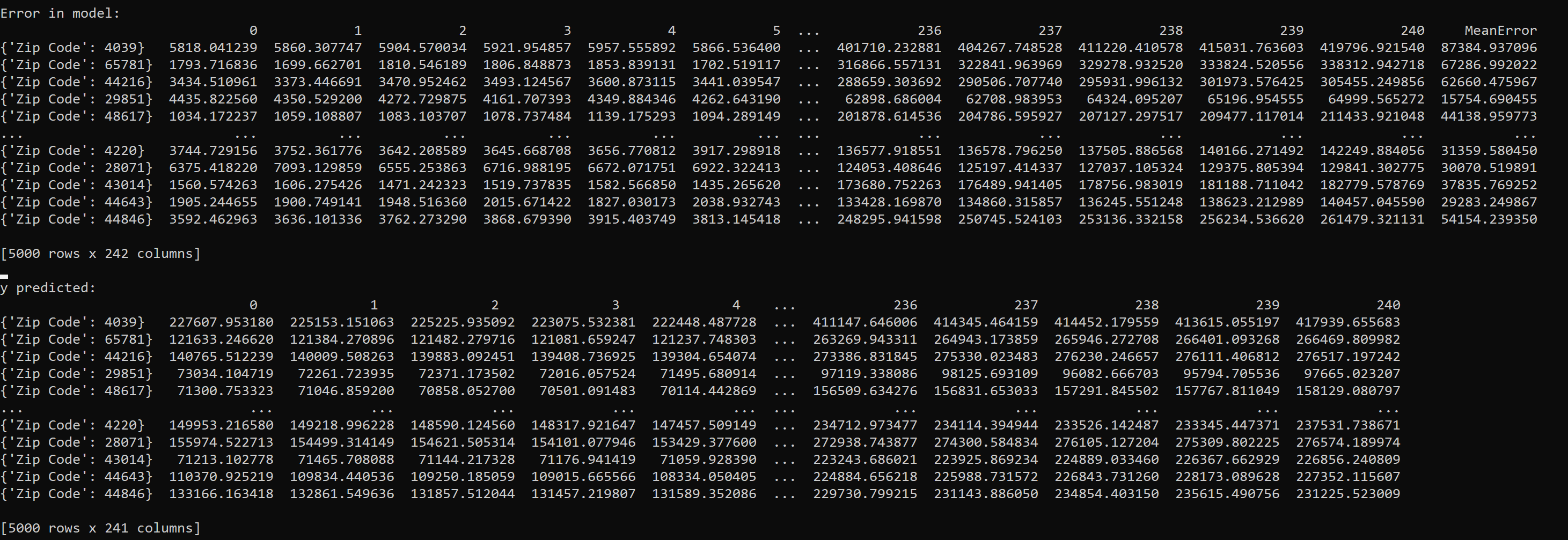
Error in model (SEA 1 yr)



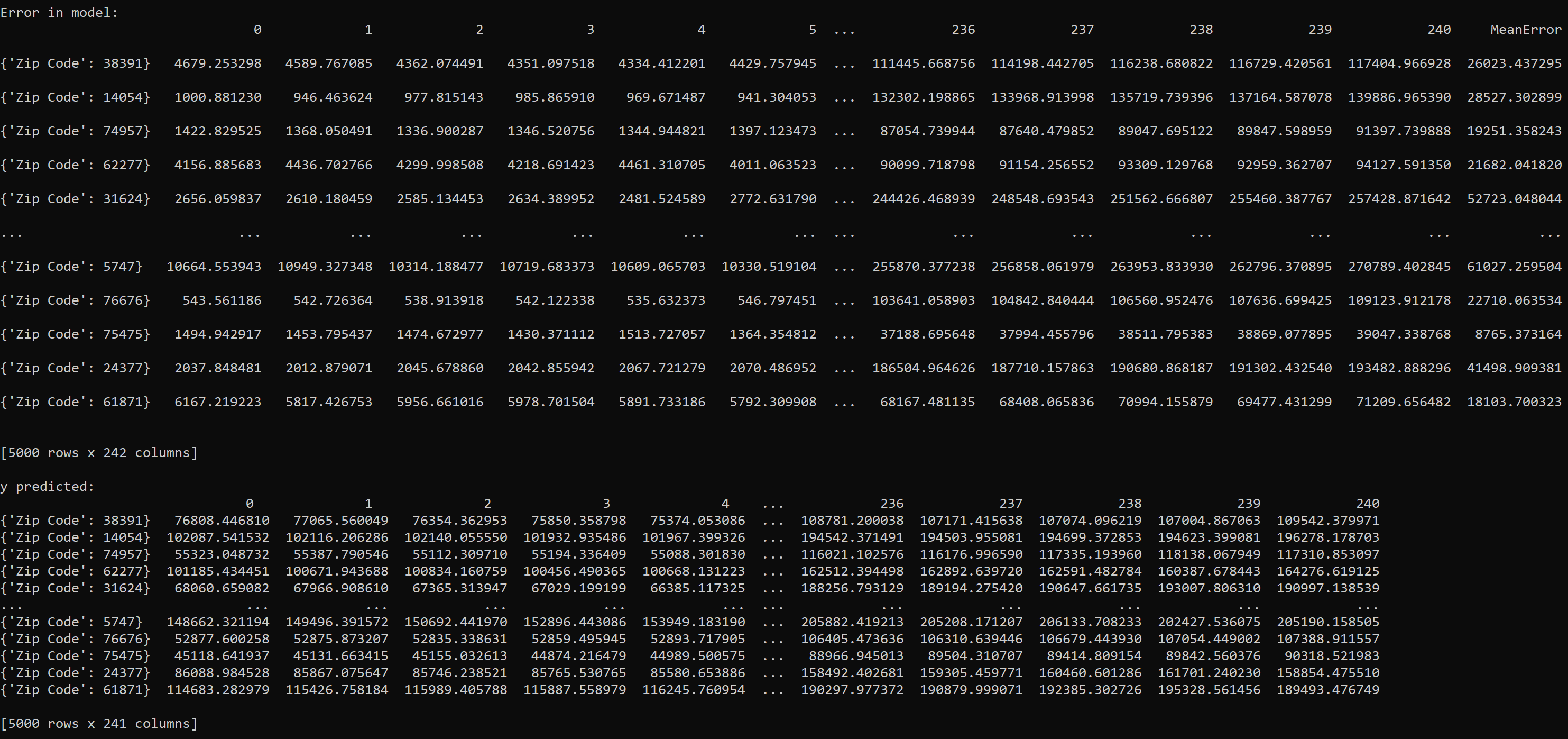
Zips 1 model output



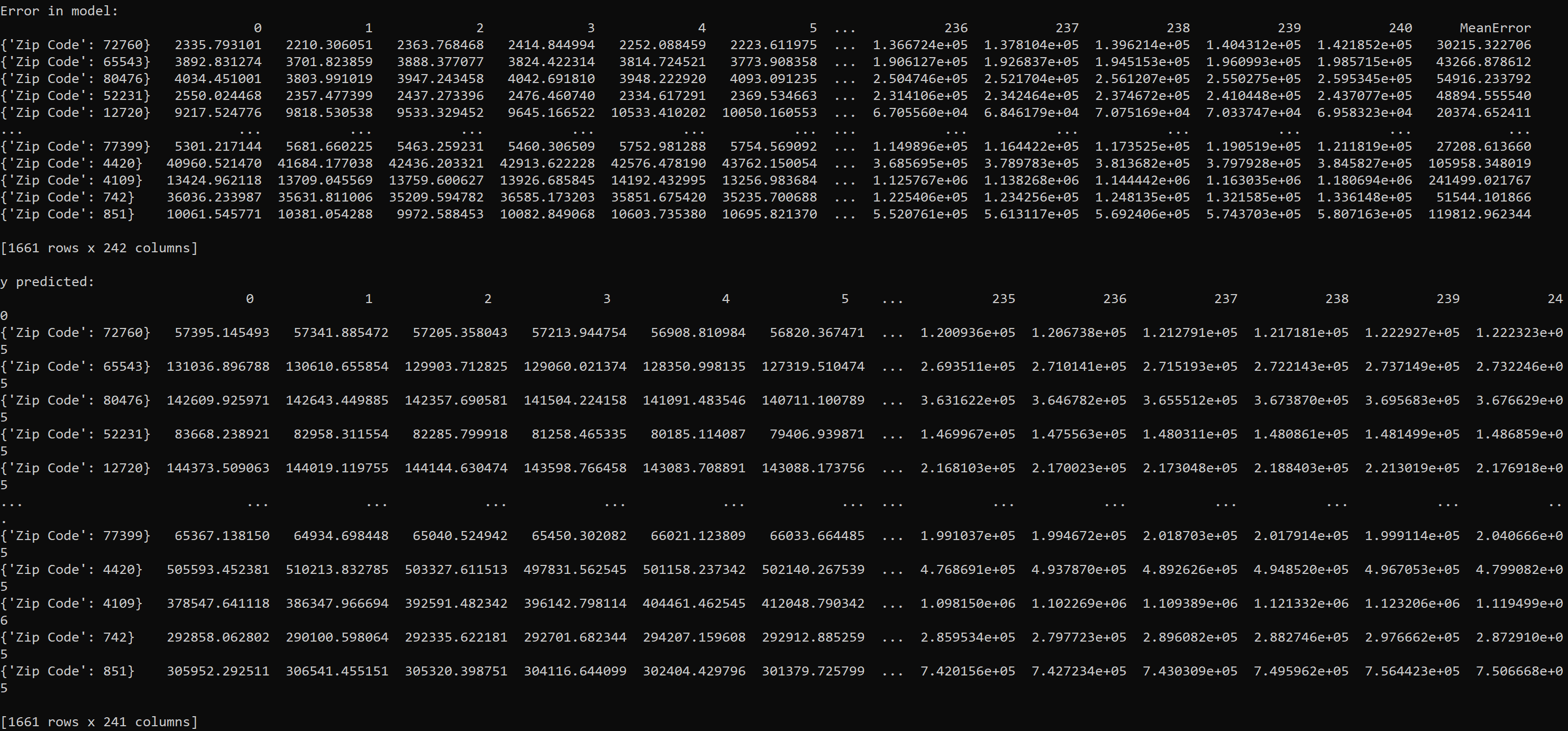
Zips 2 model output



Zips 3 model output



Zips 4 model output



Zips 5