FORECASTING REAL ESTATE PRICES USING TIME SERIES MODEL

Introduction

Laser investment firm is a reputable real estate investment firm that is dedicated to assist their clients to achieve their financial goal through implementing strategic real estate investment. It boasts of its ability to deliver exceptional value to their clients by identifying and capitalizing on lucrative real estate opportunities.

Our main goal is to forecast the top best zipcodes to invest in by use of time series model.

PROBLEM STATEMENT

Laser investment firm wants to know the top 5 best zipcodes to invest in. For this to be effective and achievable, they need to have a deep understanding on how the trends on real estate investments have been over the past years as seen in 'time-series/zillow_data.csv' dataset. Moreover, our goal is to complete this real-world task in regard to time series modeling to help answer the questions considering there could be some form of ambiguity. We will look at valuable insights by looking at the Return On Investment and Co-efficient of variation over the past years to help plan and make informed decisions.

MAIN OBJECTIVE

The main objective of this project is to design and implement a time series model that can effectively help forecast real estate prices for investments.

SPECIFIC OBJECTIVES

1.Top 5 best zip codes to invest in

2. Recommendations based on profit margin

RESEARCH QUESTIONS

1. What are the top 5 best zip codes to invest in?

2. What recommendations can you give based on profit margin?

3. Are there any risks involved in investmenting in the zipcode areas?

DATA UNDERSTANDING

For this project, we shall use 'time-series/zillow_data.cvs' dataset to analyze the real estate prices from 1996 to 2018 so as to help decide which areas to invest in.The columns in the dataset are:

1.RegionID \

2.RegionName \

3.City \

4.State \

5.Metro\

6.CountyName\

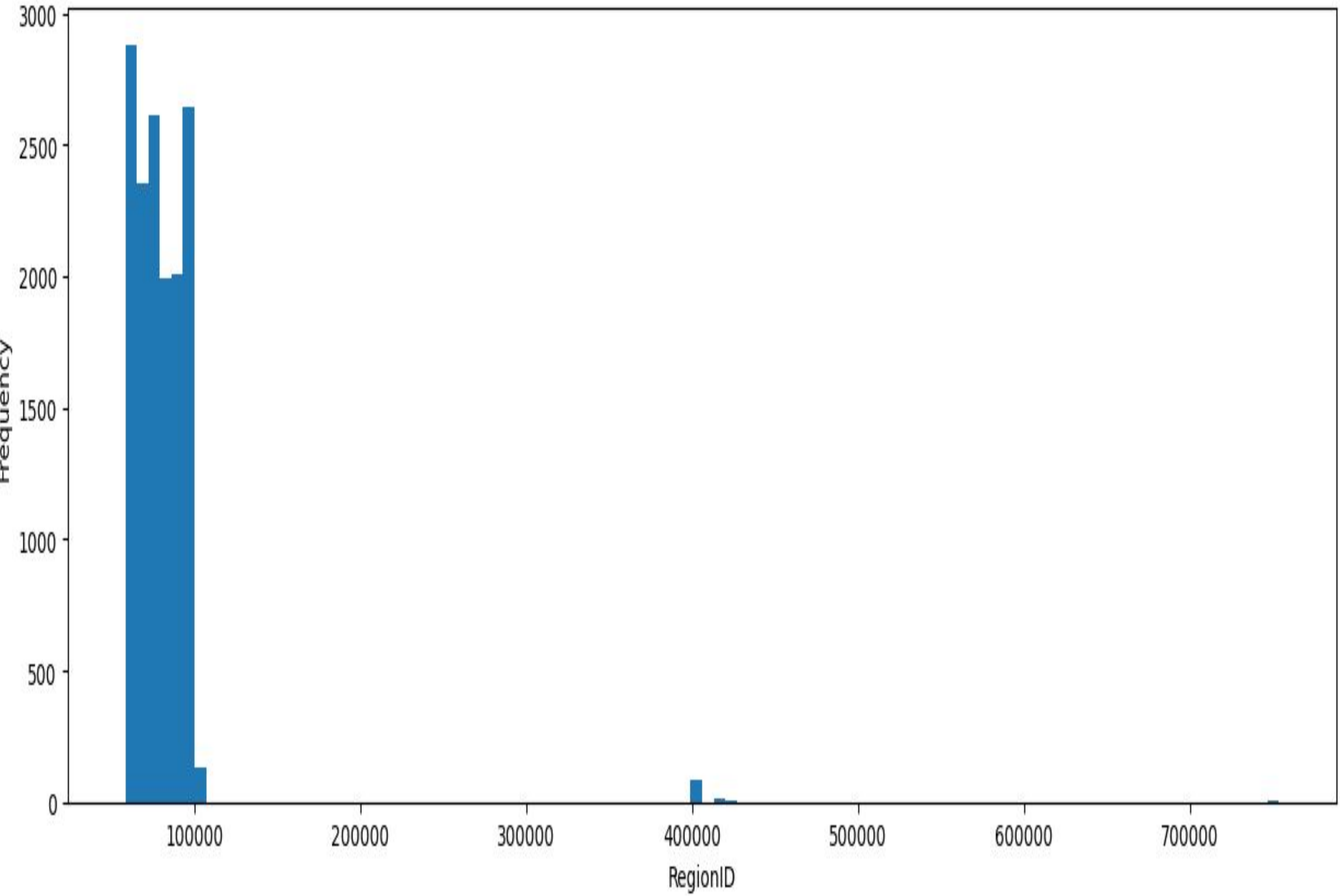
7.SizeRank \

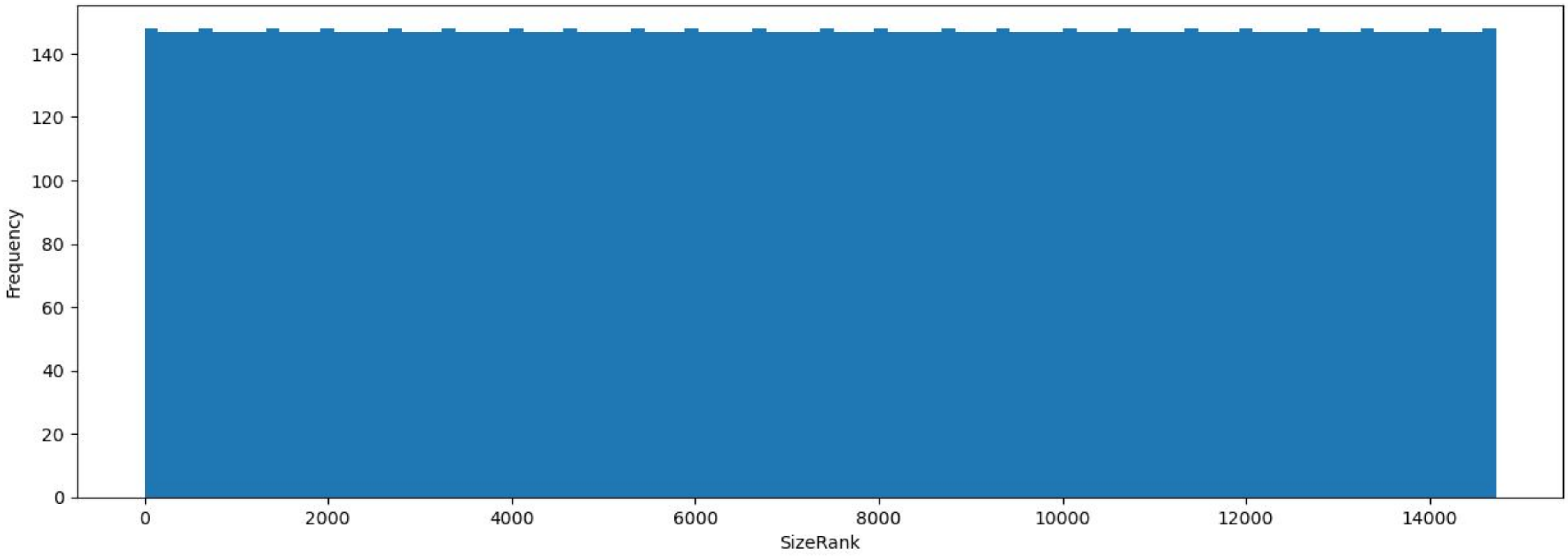
8..1996-04 to 2018-04 columns

DATA PREPARATION

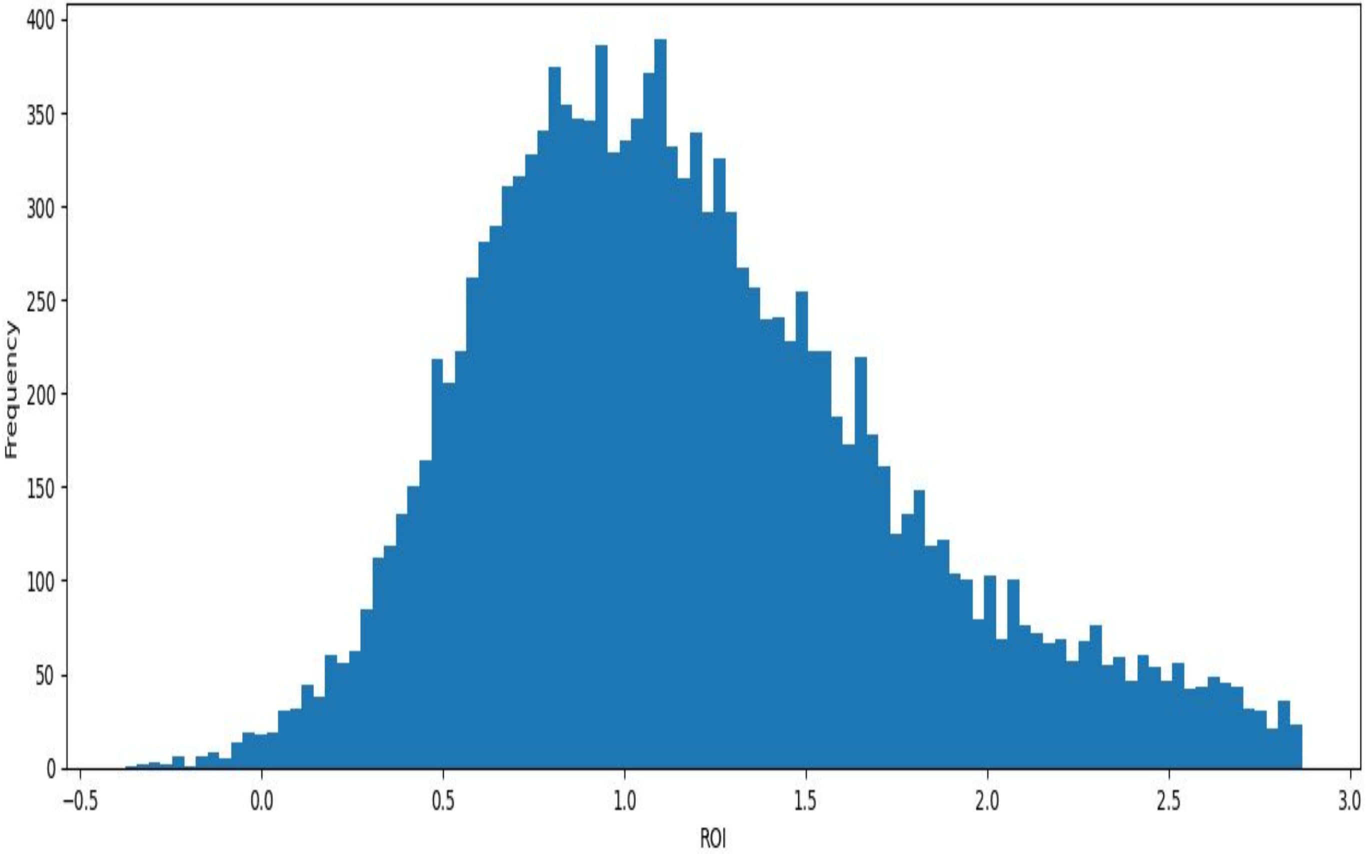
1. We loaded the data and renamed the regionName column to ZipCode.
2. We looked for outliers and handled them.
3. We looked for missing values and filled them with previous ones.
4. We preprocessed data by creating Return on Investment and CV(risk) columns.
5.

EDA AND VISUALIZATION

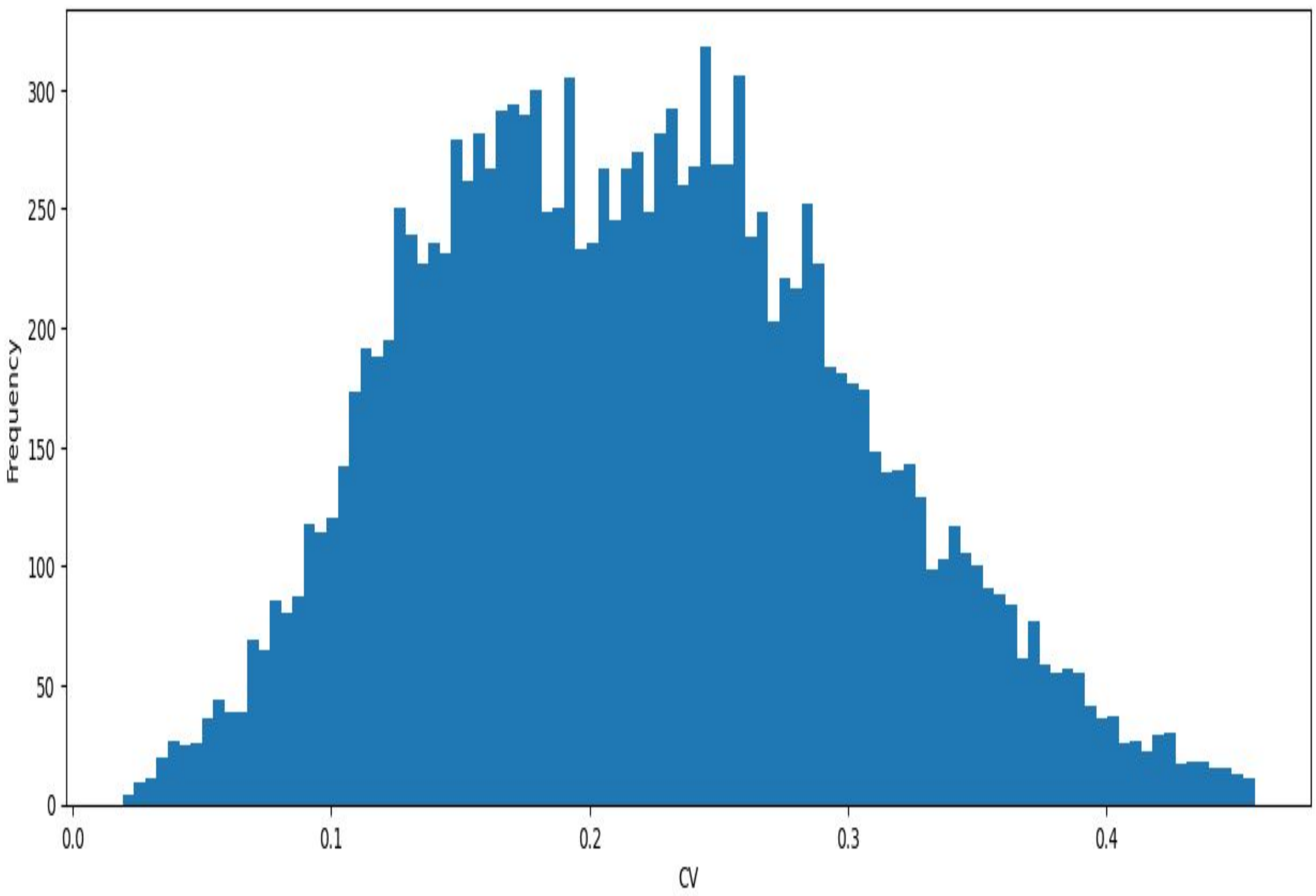




Frequency
140
120
100
80
60
40
20
0
0
2000
4000
6000
8000
10000
12000
14000
SizeRank



400
350
300
250
200
150
100
50
0
Frequency
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
ROI



Frequency
300
250
200
150
100
50
0
0.0
0.1
0.2
0.3
0.4
CV

ARMA MODEL 1

```
      SARIMAX Results
==============================================================================
Dep. Variable:                  value   No. Observations:                  264
Model:                 ARIMA(1, 0, 0)   Log Likelihood               -1942.902
Date:                Sun, 17 Sep 2023   AIC                           3891.804
Time:                        20:50:21   BIC                           3902.532
Sample:                    05-01-1996   HQIC                          3896.115
                         - 04-01-2018
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        418.5606    308.199      1.358      0.174    -185.498    1022.619
ar.L1          0.9234      0.021     43.051      0.000       0.881       0.965
sigma2      1.437e+05   9374.772     15.330      0.000    1.25e+05    1.62e+05
===================================================================================
Ljung-Box (L1) (Q):                  41.60   Jarque-Bera (JB):                61.49
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):              19.00   Skew:                             0.44
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.20
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

ARMA MODEL 2

```
SARIMAX Results
==============================================================================
Dep. Variable:                  value   No. Observations:                  264
Model:                 ARIMA(2, 0, 1)   Log Likelihood               -1900.746
Date:                Sun, 17 Sep 2023   AIC                           3811.492
Time:                        20:50:33   BIC                           3829.372
Sample:                    05-01-1996   HQIC                          3818.677
                         - 04-01-2018
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        418.5605    226.918      1.845      0.065     -26.191     863.313
ar.L1          0.9601      0.065     14.707      0.000       0.832       1.088
ar.L2         -0.1010      0.067     -1.501      0.133      -0.233       0.031
ma.L1          0.5446      0.068      8.006      0.000       0.411       0.678
sigma2      1.049e+05   6447.785     16.266      0.000    9.22e+04    1.18e+05
===============================================================================
=
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                58.39
Prob(Q):                              0.83   Prob(JB):                         0.00
Heteroskedasticity (H):              11.48   Skew:                             0.02
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.30
===============================================================================
=

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

ARMA MODEL 3

```
SARIMAX Results
==============================================================================
Dep. Variable:                  value   No. Observations:                  264
Model:                 ARIMA(2, 0, 2)   Log Likelihood               -1900.071
Date:                Sun, 17 Sep 2023   AIC                           3812.142
Time:                        20:50:46   BIC                           3833.598
Sample:                    05-01-1996   HQIC                          3820.764
                         - 04-01-2018
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        418.5569    385.948      1.084      0.278    -337.887    1175.001
ar.L1          1.6903      0.157     10.759      0.000       1.382       1.998
ar.L2         -0.7012      0.141     -4.963      0.000      -0.978      -0.424
ma.L1         -0.2732      0.160     -1.710      0.087      -0.586       0.040
ma.L2         -0.5362      0.087     -6.134      0.000      -0.708      -0.365
sigma2      1.029e+05   6542.786     15.734      0.000    9.01e+04    1.16e+05
===================================================================================
Ljung-Box (L1) (Q):                   1.41   Jarque-Bera (JB):                49.41
Prob(Q):                              0.23   Prob(JB):                         0.00
Heteroskedasticity (H):              11.96   Skew:                             0.11
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.11
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

INTERPRATING RESULTS

From the first model which is our baseline model we see that:\

The constant coefficient is not statistically significant because the p-value is 0.174, which is greater than the typical significance level of 0.05.\

The ar.L1 coefficient is highly statistically significant with a very low p-value since it is close to zero, indicating a strong positive autocorrelation in the data at lag 1.\

The sigma2 represents the estimated variance of the residuals.

From the second model, we see that:\

The constant in our ARIMA model, the estimated constant is approximately 418.5605.

The ar.L1, the estimated value is approximately 0.9601, indicating a strong positive correlation.

The ar.L2, the estimated value is approximately -0.1010, suggesting a weaker negative correlation.

The ma.L1, which reflects the impact of past white noise error terms on the current observation. Therefore, the estimated value is approximately 0.5446, indicating a positive impact from the previous error term.

The sigma2, the estimated variance is approximately 1.049e+05.

From the final model, we see that:
The constant, shows the estimated constant is approximately 418.5569.

The ar.L1, shows the estimated value is approximately 1.6903, indicating a strong positive correlation.

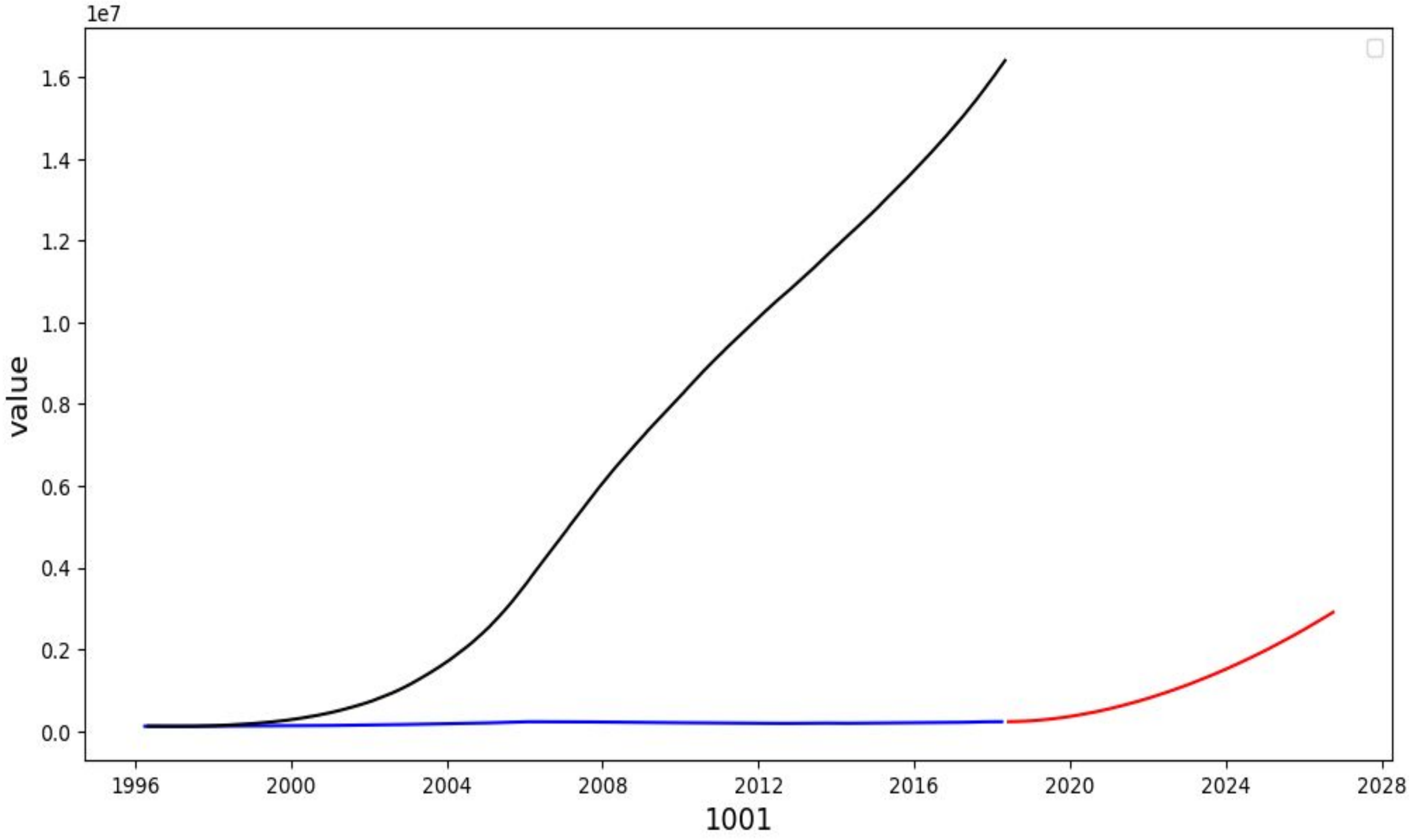
The ar.L2, shows the estimated value is approximately -0.7012, suggesting a negative correlation.

The ma.L1, shows the estimated value is approximately -0.2732, indicating a negative impact from the previous error term.

The ma.L2, shows the estimated value is approximately -0.5362, indicating a negative impact.

The sigma2, in this case, the estimated variance is approximately 1.029e+05.

PREDICTIONS



CONCLUSION

The final ARMA model after iterations achieved a lower AIC and BIC value than the baseline model(ARMA model 1) which was 3812.142 and 3833.598 respectively indicating a better model.
The Root Mean Squared Error (RMSE) of predicted 7932701.297595864 indicating that deviation from the true values.
This shows a higher prediction than the previous model.
From ROI and CV columns above, we can be able to get the top 5 zipcode to invest in due to their high Return on Investment and CV.
When making decisions, we need to balance between the risk represented as CV and return potential represented as ROI.
The higher the risk the higher the returns.

RECOMMENDATION

1. The company should consider investing in York(NewYork) which has the highest ROI of 2.8671, Santa Barbara (2.8667), Person (2.8569), Marin (2.8565), San Diego (2.8561)
2. Other factors such as infrastructure development should also be considered as it may have a big effect on house prices in the areas.
3. Cities such as Austin should be thoroughly explored due to the high price shown.