Top 5 Tech Cities worth Investing

Authors: Alice Agrawal, Jordan Kominsky, Kyongmin So, Tyler Wood, Hanis Zulmuthi

May 2022

Overview

Real estate has always been one of the most dependable markets when looking for consistent, yet high returns. Even after the housing crash of 2008, it only took a few years for the market to return to previous highs, and it has maintained steady growth ever since. This has been especially true in markets with high incomes and high density such as San Francisco or New York. Tech jobs are also gaining prominence in the job market and acquiring real estate that could serve these migrating employees could give us a competitive edge.

Business Understanding

Tech is becoming a larger part of both the US and the Global economy every year. As tech grows in a city, it doesn't only bring tech jobs, it also brings other facets of culture. In the main tech hubs of America, you'll find much more than just the industrious culture of modern technology; there will be growth in art exhibits, breweries, parks, and many other places where people can share experiences. These traits make tech cities desirable places of residence not only for those in technology, but also anyone who values being in a place that is culturally engaging.

As time goes on, less people are deciding to stay in their small towns and are moving to larger cities instead. We can see a chart from business insider that portrays the shrinking of rural America here. Furthermore, when we look at the growth of cities, we find that the largest cities are growing at the fastest rate. This article from the Brookings Institute mentions this phenomenon.

A large proportion of this growth will most likely be seen in these emerging tech hubs due to their wide cultural and employment appeal. We selected 10 cities to analyze in America that we think hold promise as places of high growth. We decided on these 10 due to an Indeed article that asserted these cities as places of high prominence in the tech industry. Specifically, these cities were Washington D.C., New York City, Seattle, San Francisco, Los Angeles, San Jose, Dallas, Boston, Chicago, and Baltimore. Many of these places have expensive markets already, but there is no shortage of demand for housing in any of these cities. As the tech sector continues to grow, there will be an even greater need to develop housing. The political landscape is starting to warm up to higher density developments such as multiplexes, which will allow for new housing development opportunities in these markets that have previously been unprofitable. Focusing on these high growth areas will provide us an advantage over the competition that is more cautious to invest in these markets with higher upfront investment barriers.

Data Understanding

Home Price

Source: Zillow Dataset

Contents: We acquired data for 14,723 different zip codes in America. The data provided monthly data on the median home price for every zip code from April 1996 to April 2018. We selected the data from the 10 prominent tech cities specified earlier, and ran a time series analysis on all of them.

2017 Median Income

Source: Kaggle Dataset

Contents: We also found data on the median income for our 10 cities. We used this to look at the home price to income ratio for our 10 cities and compare it to the U.S. average of 5.75.

Data Cleaning & Preparation

Import packages

In [2]:
```
# Basics
import pandas as pd
import numpy as np
pd.set_option('display.max_rows', 1000)
from datetime import datetime as dt

# Visualization
import matplotlib
import matplotlib.pyplot as plt
import plotly.express as px
import seaborn as sns
import matplotlib.patches as mpatches

# Modeling
from sklearn.metrics import mean_squared_error
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.stattools import adfuller
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.stattools import acf, pacf, adfuller
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from sklearn.metrics import mean_squared_error

# Warnings
import warnings
from statsmodels.tools.sm_exceptions import ConvergenceWarning
warnings.simplefilter('ignore', ConvergenceWarning)
```

Home Price

In [3]:
```
# load in zillow home price data as df
df = pd.read_csv('Data/zillow_data.csv')
df
```

Out[3]:

	RegionID	RegionName	City	State	Metro	CountyName	SizeRank	1996-04	1
0	84654	60657	Chicago	IL	Chicago	Cook	1	334200.0	3
1	90668	75070	McKinney	TX	Dallas-Fort Worth	Collin	2	235700.0	2
2	91982	77494	Katy	TX	Houston	Harris	3	210400.0	1
3	84616	60614	Chicago	IL	Chicago	Cook	4	498100.0	5
4	93144	79936	El Paso	TX	El Paso	El Paso	5	77300.0	
...	...	...	...	...	...	...	...	...	...
14718	58333	1338	Ashfield	MA	Greenfield Town	Franklin	14719	94600.0	
14719	59107	3293	Woodstock	NH	Claremont	Grafton	14720	92700.0	
14720	75672	40404	Berea	KY	Richmond	Madison	14721	57100.0	
14721	93733	81225	Mount Crested Butte	CO	NaN	Gunnison	14722	191100.0	1
14722	95851	89155	Mesquite	NV	Las Vegas	Clark	14723	176400.0	1

14723 rows × 272 columns

In [4]:
```
# we have some nulls for some of the zip codes for earlier dates, and some missi
df.info(verbose=True, null_counts=True)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14723 entries, 0 to 14722
Data columns (total 272 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   RegionID    14723 non-null  int64
 1   RegionName  14723 non-null  int64
 2   City        14723 non-null  object
 3   State       14723 non-null  object
 4   Metro       13680 non-null  object
 5   CountyName  14723 non-null  object
 6   SizeRank    14723 non-null  int64
 7   1996-04     13684 non-null  float64
 8   1996-05     13684 non-null  float64
 9   1996-06     13684 non-null  float64
 10  1996-07     13684 non-null  float64
 11  1996-08     13684 non-null  float64
 12  1996-09     13684 non-null  float64
 13  1996-10     13684 non-null  float64
 14  1996-11     13684 non-null  float64
 15  1996-12     13684 non-null  float64
 16  1997-01     13684 non-null  float64
 17  1997-02     13684 non-null  float64
 18  1997-03     13684 non-null  float64
 19  1997-04     13684 non-null  float64
 20  1997-05     13684 non-null  float64
```

```
21  1997-06     13684 non-null  float64
22  1997-07     13685 non-null  float64
23  1997-08     13685 non-null  float64
24  1997-09     13685 non-null  float64
25  1997-10     13685 non-null  float64
26  1997-11     13685 non-null  float64
27  1997-12     13685 non-null  float64
28  1998-01     13687 non-null  float64
29  1998-02     13687 non-null  float64
30  1998-03     13687 non-null  float64
31  1998-04     13687 non-null  float64
32  1998-05     13687 non-null  float64
33  1998-06     13687 non-null  float64
34  1998-07     13687 non-null  float64
35  1998-08     13687 non-null  float64
36  1998-09     13687 non-null  float64
37  1998-10     13687 non-null  float64
38  1998-11     13687 non-null  float64
39  1998-12     13687 non-null  float64
40  1999-01     13687 non-null  float64
41  1999-02     13687 non-null  float64
42  1999-03     13687 non-null  float64
43  1999-04     13687 non-null  float64
44  1999-05     13687 non-null  float64
45  1999-06     13687 non-null  float64
46  1999-07     13687 non-null  float64
47  1999-08     13687 non-null  float64
48  1999-09     13687 non-null  float64
49  1999-10     13687 non-null  float64
50  1999-11     13687 non-null  float64
51  1999-12     13687 non-null  float64
52  2000-01     13687 non-null  float64
53  2000-02     13687 non-null  float64
54  2000-03     13687 non-null  float64
55  2000-04     13687 non-null  float64
56  2000-05     13687 non-null  float64
57  2000-06     13687 non-null  float64
58  2000-07     13687 non-null  float64
59  2000-08     13687 non-null  float64
60  2000-09     13687 non-null  float64
61  2000-10     13687 non-null  float64
62  2000-11     13687 non-null  float64
63  2000-12     13687 non-null  float64
64  2001-01     13687 non-null  float64
65  2001-02     13687 non-null  float64
66  2001-03     13687 non-null  float64
67  2001-04     13687 non-null  float64
68  2001-05     13687 non-null  float64
69  2001-06     13687 non-null  float64
70  2001-07     13687 non-null  float64
71  2001-08     13687 non-null  float64
72  2001-09     13687 non-null  float64
73  2001-10     13687 non-null  float64
74  2001-11     13687 non-null  float64
75  2001-12     13687 non-null  float64
76  2002-01     13687 non-null  float64
77  2002-02     13687 non-null  float64
78  2002-03     13687 non-null  float64
79  2002-04     13687 non-null  float64
80  2002-05     13687 non-null  float64
81  2002-06     13687 non-null  float64
82  2002-07     13687 non-null  float64
83  2002-08     13687 non-null  float64
84  2002-09     13687 non-null  float64
85  2002-10     13687 non-null  float64
```

```
86  2002-11     13687 non-null  float64
87  2002-12     13687 non-null  float64
88  2003-01     13687 non-null  float64
89  2003-02     13687 non-null  float64
90  2003-03     13687 non-null  float64
91  2003-04     13687 non-null  float64
92  2003-05     13687 non-null  float64
93  2003-06     13687 non-null  float64
94  2003-07     13805 non-null  float64
95  2003-08     13805 non-null  float64
96  2003-09     13805 non-null  float64
97  2003-10     13805 non-null  float64
98  2003-11     13805 non-null  float64
99  2003-12     13805 non-null  float64
100 2004-01     13836 non-null  float64
101 2004-02     13836 non-null  float64
102 2004-03     13836 non-null  float64
103 2004-04     13836 non-null  float64
104 2004-05     13836 non-null  float64
105 2004-06     13836 non-null  float64
106 2004-07     13857 non-null  float64
107 2004-08     13857 non-null  float64
108 2004-09     13857 non-null  float64
109 2004-10     13857 non-null  float64
110 2004-11     13857 non-null  float64
111 2004-12     13857 non-null  float64
112 2005-01     13909 non-null  float64
113 2005-02     13909 non-null  float64
114 2005-03     13922 non-null  float64
115 2005-04     13922 non-null  float64
116 2005-05     13922 non-null  float64
117 2005-06     13922 non-null  float64
118 2005-07     14000 non-null  float64
119 2005-08     14000 non-null  float64
120 2005-09     14000 non-null  float64
121 2005-10     14000 non-null  float64
122 2005-11     14000 non-null  float64
123 2005-12     14000 non-null  float64
124 2006-01     14056 non-null  float64
125 2006-02     14056 non-null  float64
126 2006-03     14056 non-null  float64
127 2006-04     14056 non-null  float64
128 2006-05     14056 non-null  float64
129 2006-06     14056 non-null  float64
130 2006-07     14083 non-null  float64
131 2006-08     14083 non-null  float64
132 2006-09     14083 non-null  float64
133 2006-10     14083 non-null  float64
134 2006-11     14083 non-null  float64
135 2006-12     14083 non-null  float64
136 2007-01     14103 non-null  float64
137 2007-02     14103 non-null  float64
138 2007-03     14103 non-null  float64
139 2007-04     14103 non-null  float64
140 2007-05     14103 non-null  float64
141 2007-06     14103 non-null  float64
142 2007-07     14110 non-null  float64
143 2007-08     14110 non-null  float64
144 2007-09     14110 non-null  float64
145 2007-10     14110 non-null  float64
146 2007-11     14110 non-null  float64
147 2007-12     14110 non-null  float64
148 2008-01     14116 non-null  float64
149 2008-02     14116 non-null  float64
150 2008-03     14116 non-null  float64
```

```
151 2008-04     14116 non-null  float64
152 2008-05     14116 non-null  float64
153 2008-06     14116 non-null  float64
154 2008-07     14125 non-null  float64
155 2008-08     14125 non-null  float64
156 2008-09     14125 non-null  float64
157 2008-10     14125 non-null  float64
158 2008-11     14125 non-null  float64
159 2008-12     14125 non-null  float64
160 2009-01     14136 non-null  float64
161 2009-02     14136 non-null  float64
162 2009-03     14136 non-null  float64
163 2009-04     14136 non-null  float64
164 2009-05     14136 non-null  float64
165 2009-06     14136 non-null  float64
166 2009-07     14143 non-null  float64
167 2009-08     14143 non-null  float64
168 2009-09     14143 non-null  float64
169 2009-10     14143 non-null  float64
170 2009-11     14143 non-null  float64
171 2009-12     14143 non-null  float64
172 2010-01     14144 non-null  float64
173 2010-02     14144 non-null  float64
174 2010-03     14374 non-null  float64
175 2010-04     14374 non-null  float64
176 2010-05     14374 non-null  float64
177 2010-06     14374 non-null  float64
178 2010-07     14415 non-null  float64
179 2010-08     14415 non-null  float64
180 2010-09     14415 non-null  float64
181 2010-10     14415 non-null  float64
182 2010-11     14415 non-null  float64
183 2010-12     14415 non-null  float64
184 2011-01     14448 non-null  float64
185 2011-02     14448 non-null  float64
186 2011-03     14448 non-null  float64
187 2011-04     14448 non-null  float64
188 2011-05     14448 non-null  float64
189 2011-06     14448 non-null  float64
190 2011-07     14472 non-null  float64
191 2011-08     14472 non-null  float64
192 2011-09     14472 non-null  float64
193 2011-10     14472 non-null  float64
194 2011-11     14472 non-null  float64
195 2011-12     14472 non-null  float64
196 2012-01     14499 non-null  float64
197 2012-02     14499 non-null  float64
198 2012-03     14499 non-null  float64
199 2012-04     14499 non-null  float64
200 2012-05     14499 non-null  float64
201 2012-06     14499 non-null  float64
202 2012-07     14517 non-null  float64
203 2012-08     14517 non-null  float64
204 2012-09     14517 non-null  float64
205 2012-10     14517 non-null  float64
206 2012-11     14517 non-null  float64
207 2012-12     14517 non-null  float64
208 2013-01     14572 non-null  float64
209 2013-02     14572 non-null  float64
210 2013-03     14572 non-null  float64
211 2013-04     14572 non-null  float64
212 2013-05     14572 non-null  float64
213 2013-06     14572 non-null  float64
214 2013-07     14614 non-null  float64
215 2013-08     14614 non-null  float64
```

```
 216 2013-09     14614 non-null  float64
 217 2013-10     14614 non-null  float64
 218 2013-11     14614 non-null  float64
 219 2013-12     14614 non-null  float64
 220 2014-01     14667 non-null  float64
 221 2014-02     14667 non-null  float64
 222 2014-03     14667 non-null  float64
 223 2014-04     14667 non-null  float64
 224 2014-05     14667 non-null  float64
 225 2014-06     14667 non-null  float64
 226 2014-07     14723 non-null  int64
 227 2014-08     14723 non-null  int64
 228 2014-09     14723 non-null  int64
 229 2014-10     14723 non-null  int64
 230 2014-11     14723 non-null  int64
 231 2014-12     14723 non-null  int64
 232 2015-01     14723 non-null  int64
 233 2015-02     14723 non-null  int64
 234 2015-03     14723 non-null  int64
 235 2015-04     14723 non-null  int64
 236 2015-05     14723 non-null  int64
 237 2015-06     14723 non-null  int64
 238 2015-07     14723 non-null  int64
 239 2015-08     14723 non-null  int64
 240 2015-09     14723 non-null  int64
 241 2015-10     14723 non-null  int64
 242 2015-11     14723 non-null  int64
 243 2015-12     14723 non-null  int64
 244 2016-01     14723 non-null  int64
 245 2016-02     14723 non-null  int64
 246 2016-03     14723 non-null  int64
 247 2016-04     14723 non-null  int64
 248 2016-05     14723 non-null  int64
 249 2016-06     14723 non-null  int64
 250 2016-07     14723 non-null  int64
 251 2016-08     14723 non-null  int64
 252 2016-09     14723 non-null  int64
 253 2016-10     14723 non-null  int64
 254 2016-11     14723 non-null  int64
 255 2016-12     14723 non-null  int64
 256 2017-01     14723 non-null  int64
 257 2017-02     14723 non-null  int64
 258 2017-03     14723 non-null  int64
 259 2017-04     14723 non-null  int64
 260 2017-05     14723 non-null  int64
 261 2017-06     14723 non-null  int64
 262 2017-07     14723 non-null  int64
 263 2017-08     14723 non-null  int64
 264 2017-09     14723 non-null  int64
 265 2017-10     14723 non-null  int64
 266 2017-11     14723 non-null  int64
 267 2017-12     14723 non-null  int64
 268 2018-01     14723 non-null  int64
 269 2018-02     14723 non-null  int64
 270 2018-03     14723 non-null  int64
 271 2018-04     14723 non-null  int64
dtypes: float64(219), int64(49), object(4)
memory usage: 30.6+ MB
```

In [5]:
```
# The min of the zipcodes is only 4 digits, this is because some zips start with
df.describe()
```

Out[5]:

	RegionID	RegionName	SizeRank	1996-04	1996-05	1996-06

	RegionID	RegionName	SizeRank	1996-04	1996-05	1996-06
count	14723.000000	14723.000000	14723.000000	1.368400e+04	1.368400e+04	1.368400e+04
mean	81075.010052	48222.348706	7362.000000	1.182991e+05	1.184190e+05	1.185374e+05
std	31934.118525	29359.325439	4250.308342	8.600251e+04	8.615567e+04	8.630923e+04
min	58196.000000	1001.000000	1.000000	1.130000e+04	1.150000e+04	1.160000e+04
25%	67174.500000	22101.500000	3681.500000	6.880000e+04	6.890000e+04	6.910000e+04
50%	78007.000000	46106.000000	7362.000000	9.950000e+04	9.950000e+04	9.970000e+04
75%	90920.500000	75205.500000	11042.500000	1.432000e+05	1.433000e+05	1.432250e+05
max	753844.000000	99901.000000	14723.000000	3.676700e+06	3.704200e+06	3.729600e+06

8 rows × 268 columns

In [6]:
```
# for example, Agawam Massachusetts zip code is 01001
df[df['RegionName'] == 1001]
```

Out[6]:

	RegionID	RegionName	City	State	Metro	CountyName	SizeRank	1996-04	19
5850	58196	1001	Agawam	MA	Springfield	Hampden	5851	113100.0	1128(

1 rows × 272 columns

Many cities may share names with cities in other states. For example, there are 13 different cities named Washington in America. In order to quickly select only the cities that we want, we created a column that includes both the city and the state as a string.

In [7]:
```
# list of all the different states with a city named Washington
df[df['City'] == 'Washington']['State'].unique()
```

Out[7]:
```
array(['DC', 'PA', 'NJ', 'IL', 'MO', 'UT', 'MI', 'IN', 'IA', 'WV', 'NY',
       'CT', 'NH'], dtype=object)
```

In [8]:
```
# combines the City and State columns as a new Geolocate columns
df['Geolocate'] = df['City'] + ', ' + df['State']
```

Mask data to cities of interest

We're interested in looking at the housing market of major tech cities in the US. We know that these markets are prominent in the technology sector, and limiting our model to only these large cities will allow us to create a better model th at fits large city real estate dynamics.

In [9]:
```
# list of the 10 tech cities
city_list = ['Washington, DC','New York, NY','San Francisco, CA','Seattle, WA',\
             'Dallas, TX','Los Angeles, CA','San Jose, CA','Chicago, IL','Baltim
```

In [10]:
```
# create a df that only contains our 10
df_cities = df[df.Geolocate.isin(city_list)]
df_cities.head()
```

Out[10]:

	RegionID	RegionName	City	State	Metro	CountyName	SizeRank	1996-04	1996-0
0	84654	60657	Chicago	IL	Chicago	Cook	1	334200.0	335400.
3	84616	60614	Chicago	IL	Chicago	Cook	4	498100.0	500900.
6	61807	10467	New York	NY	New York	Bronx	7	152900.0	152700.
7	84640	60640	Chicago	IL	Chicago	Cook	8	216500.0	216700.
9	97564	94109	San Francisco	CA	San Francisco	San Francisco	10	766000.0	771100.

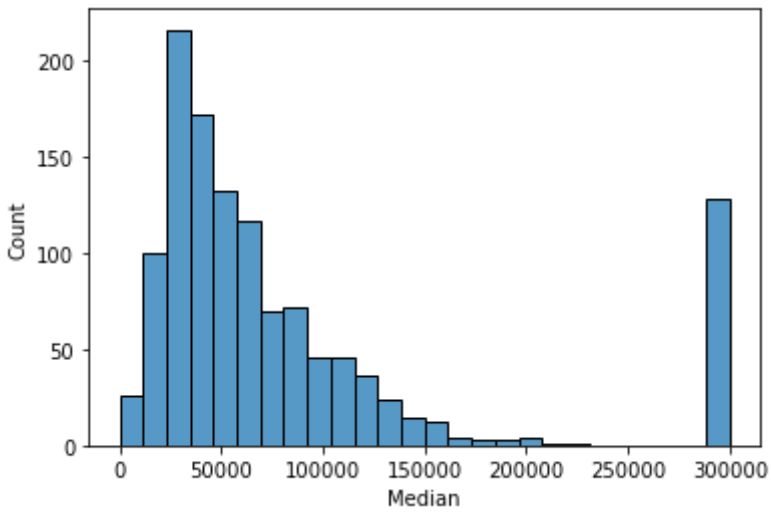
5 rows × 273 columns

2017 Median Income data

Here we thought it would be productive to look at the median income at these major tech cities. This 2017 median income data was found on Kaggle in the form of a csv. Cities with higher median incomes would be much more likely to invest in properties. However, if a cities home price to income ratio gets too high, it can end up making people more likely to lease apartments rather than buy long-term housing.

In [11]:
```
#Load income data
df_income = pd.read_csv('Data/kaggle_income.csv',
                        encoding = 'ISO-8859-1')

#Combine city and state
df_income['Geolocate'] = df_income['City'] + ', ' + df_income['State_ab']

#Mask dataframe to cities of interest
df_income = df_income[df_income.Geolocate.isin(city_list)]
```

In [12]:
```
#Look at distribution of 2017 median income
sns.histplot(df_income['Median']);
```



In [13]:
```
# summary statistics of the income df, some median incomes are 0 or 300,000
df_income.describe()
```

Out[13]:

	id	State_Code	Zip_Code	ALand	AWater	Lat	
count	1.227000e+03	1227.000000	1227.000000	1.227000e+03	1.227000e+03	1227.000000	122
mean	6.428769e+07	21.402608	60256.520782	1.881765e+06	3.313473e+05	38.731672	-9
std	1.032249e+08	15.223327	31795.355730	1.750294e+07	6.917771e+06	3.876569	
min	1.101200e+04	6.000000	2111.000000	0.000000e+00	0.000000e+00	29.659982	-12
25%	1.702394e+07	6.000000	21221.000000	3.307440e+05	0.000000e+00	34.088205	-1
50%	2.402528e+07	17.000000	60647.000000	6.745220e+05	0.000000e+00	39.302056	-8
75%	6.021341e+07	36.000000	90034.000000	1.337892e+06	0.000000e+00	41.781700	-7
max	4.802131e+08	53.000000	98199.000000	5.888084e+08	2.396078e+08	47.730188	

The distribution of the median income in the major tech cities is skewed to right. There is a little cleaning that needs to be done for this data, which could lead to a little inconsistency with actual real world figures. Interesting that the median of the 2017 median income is ~$50,000. Looking at the distribution above and the data frame below, it seems that income is higher than average in these cities, but the wealth is concentrated in specific neighborhoods. This could provide opportunities for investment in neighborhoods that could be raised in value. However, this should be done in a way that does not disrupt affordable housing and lead to gentrification.

In [14]:
```
# This will remove the zip codes equal to 0 or 300,000
df_income = df_income[df_income['Median'] != 0][df_income['Median'] != 300000]
```

```
<ipython-input-14-bbacd809ea29>:2: UserWarning: Boolean Series key will be reind
exed to match DataFrame index.
  df_income = df_income[df_income['Median'] != 0][df_income['Median'] != 300000]
```

In [15]:
```
# look at the median of median income
df_income['Median'].median()
```

Out[15]:
```
49318.5
```

In [16]:
```
# Group by cities and aggregate the values
df_income = df_income.groupby(['Geolocate']).agg({'Mean': 'mean',
                                                  'Median': 'mean',
                                                  'Stdev': 'mean'}).reset_index(
df_income[['Mean', 'Median', 'Stdev']] = df_income[['Mean', 'Median', 'Stdev']].

# Rename columns
df_income = df_income.rename(columns = {'Mean':'Mean_2017income','Median':'Media
df_income
```

Out[16]:

	Geolocate	Mean_2017income	Median_2017income	Stdev_2017income
0	Baltimore, MD	62451	53203	45935
1	Boston, MA	79133	59771	67582
2	Chicago, IL	63156	50451	49568

	Geolocate	Mean_2017income	Median_2017income	Stdev_2017income
3	Dallas, TX	65304	55669	47353
4	Los Angeles, CA	57626	45759	46520
5	New York, NY	92179	80079	66525
6	San Francisco, CA	98358	83966	72248
7	San Jose, CA	94888	84591	62358
8	Seattle, WA	82974	68733	59768
9	Washington, DC	85698	73142	62392

In [17]:
```
# Pull out 2017 home price
home_price_2017 = df_cities.groupby('Geolocate').median()
home_price_2017 = home_price_2017[['2017-12']]
home_price_2017 = home_price_2017.reset_index()
home_price_2017
```

Out[17]:

	Geolocate	2017-12
0	Baltimore, MD	160450.0
1	Boston, MA	551200.0
2	Chicago, IL	355200.0
3	Dallas, TX	326700.0
4	Los Angeles, CA	730500.0
5	New York, NY	727350.0
6	San Francisco, CA	1704500.0
7	San Jose, CA	1077100.0
8	Seattle, WA	827100.0
9	Washington, DC	771150.0

The median income in the US in 2017 was $61,000, so it's likely that this data is slightly off. Still, half of the cities have a median income well above average. The incomes are probably underestimated, so the home-price income ratios will probably be slightly inflated. Even after we qualify the observations from this graph, we can see several of the cities have housing prices over the us average ratio of 5.75 (in 2017). While these high ratios do show that many people will not be able to buy homes, they also show that there is high demand for more housing in many of these cities.

In [18]:
```
# Merge dataframes containing 2017 median income and 2017 median house price.

city_df = df_income.merge(home_price_2017)
city_df = city_df.rename(columns = {'2017-12':'Median_house_price_2017'})
city_df ['Home/income Ratio'] = city_df['Median_house_price_2017']/city_df['Medi
city_df
```

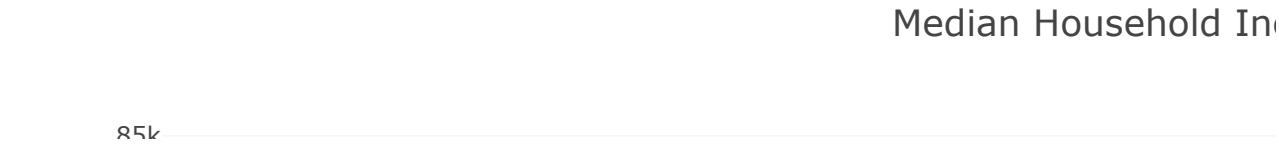
Out[18]:

	Geolocate	Mean_2017income	Median_2017income	Stdev_2017income	Median_house_price_2

	Geolocate	Mean_2017income	Median_2017income	Stdev_2017income	Median_house_price_2
0	Baltimore, MD	62451	53203	45935	1604
1	Boston, MA	79133	59771	67582	5512
2	Chicago, IL	63156	50451	49568	3552
3	Dallas, TX	65304	55669	47353	3267
4	Los Angeles, CA	57626	45759	46520	7305
5	New York, NY	92179	80079	66525	7273
6	San Francisco, CA	98358	83966	72248	17045
7	San Jose, CA	94888	84591	62358	10771
8	Seattle, WA	82974	68733	59768	8271
9	Washington, DC	85698	73142	62392	7711

Plotting all our dataframe

In [19]:
```
fig = px.scatter(city_df, x = 'Median_house_price_2017',
                 y = 'Median_2017income',
                 text="Geolocate", color = 'Median_house_price_2017', size='Home
                 size_max=25,
                 color_continuous_scale = 'Bluered', hover_name = 'Geolocate',
                 hover_data = {'Median_2017income': ':$,2f',
                               'Median_house_price_2017': ':$,2f'},
                labels={"Median_house_price_2017": "2017 Median Home price", "Me
                width=1600, height=800)

fig.update_traces(textposition=["top center", "top right", "top right","top righ
                  "top right", "top right", "middle left", "top center", "top ri

fig.update_layout({"plot_bgcolor": "rgba(0, 0, 0, 0)",
                   "paper_bgcolor": "rgba(0, 0, 0, 0)"},
                  title_text = 'Median Household Income vs. Median Home Price ($
                  title_font_size = 18,
                  title_xref = 'container',
                  title_y = 0.95,
                  title_x = 0.5,
                  showlegend = False,
                  hovermode = 'closest',
                  template = 'xgridoff')
```



Time Series Data

Prep time series data

In [20]:
```
# This function is provided with the starter notebook, changes df from wide to l
def melt_data(df,city):
    """
    Takes the zillow_data dataset in wide form or a subset of the zillow_dataset
    Returns a long-form datetime dataframe
    with the datetime column names as the index and the values as the 'values' c

    If more than one row is passes in the wide-form dataset, the values column
    will be the mean of the values from the datetime columns in all of the rows.
```

```
    """
    melted = pd.melt(df, id_vars=['RegionName', 'RegionID', 'SizeRank', 'City',
    melted['time'] = pd.to_datetime(melted['time'], infer_datetime_format=True)
    melted = melted.dropna(subset=['value'])
    melted_df= melted.groupby('time').aggregate({'value':'median'})
    melted_df.rename(columns = {'value':city}, inplace = True)
    return melted_df
```

In [21]:

```
# instantiate melted_df as an empty dataframe
melted_df = pd.DataFrame()
# run a for loop over every city in our list
for city in city_list:
    # get all the observations from the cities we need
    city_df = df[df['Geolocate'] == city]
    # use the melt function to change the format for our df
    city_melt = melt_data(city_df,city)
    # replaces the empty dataframe with city_melt for the first iteration
    if len(melted_df) == 0:
        melted_df = city_melt
    # joins city_melt with the melted_df for every subsequent iteration
    else:
        melted_df=melted_df.join(city_melt)
melted_df
```

Out[21]:

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
1996-04-01	126500.0	175950.0	306900.0	170600.0	165350.0	165500.0	234600.0	149750.0
1996-05-01	126250.0	175650.0	307600.0	171000.0	166250.0	166300.0	235100.0	149700.0
1996-06-01	126000.0	175800.0	308400.0	171600.0	166850.0	166900.0	235900.0	149450.0
1996-07-01	125800.0	175150.0	309300.0	172200.0	167100.0	166500.0	237300.0	149100.0
1996-08-01	125750.0	174400.0	310500.0	173000.0	167250.0	166200.0	238800.0	149050.0
1996-09-01	125900.0	174000.0	312000.0	173800.0	167350.0	165900.0	240400.0	148650.0
1996-10-01	126250.0	174050.0	313700.0	174800.0	167400.0	165700.0	241600.0	148000.0
1996-11-01	126650.0	173850.0	315600.0	176000.0	167350.0	165700.0	243200.0	147000.0
1996-12-01	127300.0	174000.0	318100.0	177400.0	167500.0	166000.0	245300.0	146400.0
1997-01-01	128050.0	174150.0	321000.0	179000.0	167750.0	166600.0	247700.0	145850.0
1997-02-01	128800.0	174200.0	323900.0	180600.0	167850.0	167300.0	250100.0	145100.0
1997-03-01	129500.0	174500.0	326600.0	182000.0	167800.0	167900.0	252400.0	144050.0

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
1997-04-01	130200.0	174950.0	329600.0	183600.0	168000.0	168700.0	255000.0	143150.0
1997-05-01	131050.0	175550.0	332600.0	185100.0	168450.0	169700.0	257800.0	142500.0
1997-06-01	131900.0	176400.0	335800.0	186700.0	169150.0	170800.0	260700.0	142100.0
1997-07-01	132650.0	177100.0	339200.0	188300.0	170050.0	172000.0	264400.0	141800.0
1997-08-01	133350.0	177900.0	342700.0	190000.0	171100.0	173200.0	268200.0	141850.0
1997-09-01	134000.0	178800.0	346400.0	192800.0	172200.0	174500.0	272000.0	142350.0
1997-10-01	134800.0	179800.0	350400.0	196000.0	173250.0	175900.0	275700.0	143150.0
1997-11-01	135700.0	181000.0	354500.0	199400.0	174050.0	177600.0	279400.0	144350.0
1997-12-01	136650.0	181850.0	359100.0	203100.0	174650.0	180000.0	283200.0	146150.0
1998-01-01	137600.0	182850.0	364200.0	206900.0	175100.0	182200.0	287100.0	148450.0
1998-02-01	138500.0	183700.0	369300.0	210600.0	175150.0	183900.0	290800.0	150800.0
1998-03-01	139150.0	184350.0	374100.0	214100.0	174750.0	185300.0	293900.0	153100.0
1998-04-01	139950.0	185050.0	379100.0	217600.0	174150.0	186700.0	297200.0	155600.0
1998-05-01	140750.0	185650.0	384000.0	221000.0	173400.0	188200.0	300900.0	158150.0
1998-06-01	141700.0	186350.0	388900.0	224300.0	172500.0	189900.0	304300.0	160750.0
1998-07-01	142650.0	187100.0	393700.0	227500.0	171450.0	191500.0	307000.0	163250.0
1998-08-01	143650.0	187950.0	398300.0	230700.0	170400.0	193200.0	308800.0	165800.0
1998-09-01	144750.0	189400.0	402600.0	233800.0	169300.0	194900.0	310600.0	168150.0
1998-10-01	145950.0	190500.0	406800.0	236900.0	168250.0	196500.0	312600.0	170500.0
1998-11-01	147250.0	191950.0	410900.0	240000.0	167450.0	198600.0	314900.0	172650.0
1998-12-01	148800.0	193650.0	415400.0	243100.0	166950.0	201400.0	317800.0	174700.0

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
1999-01-01	150500.0	195350.0	420200.0	246300.0	166800.0	204500.0	321200.0	176550.0
1999-02-01	152150.0	196950.0	425100.0	249200.0	166800.0	206600.0	324700.0	177950.0
1999-03-01	153650.0	198450.0	430000.0	251900.0	166900.0	207900.0	328300.0	179250.0
1999-04-01	155200.0	200000.0	435400.0	254500.0	167250.0	209200.0	332700.0	180400.0
1999-05-01	156750.0	202000.0	441300.0	257100.0	167900.0	210300.0	336700.0	181550.0
1999-06-01	158350.0	204250.0	447900.0	259700.0	168700.0	211300.0	341700.0	182600.0
1999-07-01	159950.0	206500.0	455500.0	262300.0	169600.0	212100.0	347800.0	183800.0
1999-08-01	161550.0	208850.0	463900.0	265000.0	170500.0	212900.0	354000.0	185050.0
1999-09-01	163150.0	211200.0	473100.0	267300.0	171400.0	213900.0	360800.0	186500.0
1999-10-01	164750.0	213650.0	483300.0	269600.0	172150.0	215900.0	368100.0	188150.0
1999-11-01	166350.0	216150.0	494300.0	272200.0	172800.0	218200.0	376300.0	190050.0
1999-12-01	168150.0	218900.0	506200.0	274900.0	173450.0	220600.0	384900.0	192350.0
2000-01-01	170100.0	221800.0	518700.0	277700.0	173900.0	223300.0	395000.0	195000.0
2000-02-01	171950.0	224700.0	530800.0	280500.0	174200.0	225900.0	405300.0	197650.0
2000-03-01	173650.0	227550.0	542100.0	283100.0	174300.0	228200.0	415200.0	200300.0
2000-04-01	175300.0	229750.0	552700.0	285600.0	174450.0	230400.0	424800.0	203100.0
2000-05-01	176950.0	232650.0	562300.0	288000.0	174650.0	232600.0	434000.0	206000.0
2000-06-01	178750.0	235600.0	570800.0	290400.0	175000.0	234700.0	442500.0	209050.0
2000-07-01	180650.0	238500.0	579600.0	292800.0	175550.0	236800.0	449600.0	212100.0
2000-08-01	182750.0	241250.0	588300.0	295100.0	176150.0	238900.0	455600.0	215200.0
2000-09-01	185050.0	244500.0	596100.0	297400.0	176850.0	240900.0	461100.0	217650.0

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
2000-10-01	187650.0	247800.0	603100.0	299000.0	177550.0	243200.0	466500.0	219400.0
2000-11-01	190600.0	250900.0	609300.0	300200.0	178100.0	245500.0	469800.0	221400.0
2000-12-01	194000.0	253800.0	614800.0	301600.0	178550.0	248100.0	472100.0	223850.0
2001-01-01	197650.0	256500.0	619500.0	303100.0	178950.0	250900.0	474800.0	226250.0
2001-02-01	201500.0	259000.0	623400.0	304700.0	179100.0	253800.0	477600.0	228400.0
2001-03-01	205350.0	261300.0	626400.0	306100.0	179050.0	256500.0	477700.0	230200.0
2001-04-01	209400.0	263350.0	628600.0	307400.0	178950.0	259200.0	476800.0	231950.0
2001-05-01	213550.0	265350.0	630000.0	308600.0	178950.0	261800.0	475200.0	233550.0
2001-06-01	217850.0	267250.0	631100.0	309700.0	178950.0	264200.0	473100.0	235100.0
2001-07-01	222250.0	269050.0	633000.0	310700.0	179100.0	266300.0	471800.0	236650.0
2001-08-01	226750.0	270800.0	634300.0	311300.0	179350.0	268300.0	471400.0	238200.0
2001-09-01	231400.0	272550.0	631600.0	311800.0	179600.0	270300.0	471200.0	239750.0
2001-10-01	236200.0	274400.0	629600.0	312700.0	179950.0	272500.0	471000.0	241350.0
2001-11-01	241050.0	276300.0	628800.0	314000.0	180350.0	274700.0	471100.0	242850.0
2001-12-01	245950.0	278300.0	627900.0	315600.0	180750.0	279400.0	471600.0	244450.0
2002-01-01	250800.0	280500.0	627200.0	317500.0	181150.0	284900.0	472800.0	246100.0
2002-02-01	255500.0	282800.0	628800.0	319500.0	181450.0	289900.0	474700.0	247700.0
2002-03-01	259900.0	285150.0	631600.0	321500.0	181800.0	293800.0	476200.0	249100.0
2002-04-01	264050.0	287750.0	635300.0	323300.0	182200.0	297700.0	476500.0	250550.0
2002-05-01	268000.0	290000.0	636800.0	325000.0	182650.0	301500.0	477100.0	251900.0
2002-06-01	271850.0	292150.0	637700.0	326200.0	183150.0	305300.0	477900.0	253250.0

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
2002-07-01	275350.0	294050.0	640000.0	327000.0	183700.0	309000.0	478800.0	254650.0
2002-08-01	277650.0	295950.0	646200.0	327900.0	184250.0	312600.0	479800.0	256000.0
2002-09-01	279850.0	298400.0	652500.0	328800.0	184700.0	316400.0	482500.0	257450.0
2002-10-01	282100.0	301500.0	658600.0	329200.0	184950.0	320700.0	485000.0	259150.0
2002-11-01	285100.0	303200.0	665000.0	329500.0	185150.0	327600.0	487100.0	260900.0
2002-12-01	288150.0	303250.0	670400.0	329800.0	185200.0	334800.0	488500.0	263000.0
2003-01-01	290550.0	308150.0	674200.0	330300.0	185000.0	341600.0	489900.0	265300.0
2003-02-01	292400.0	319700.0	677500.0	331100.0	184650.0	347700.0	490200.0	267400.0
2003-03-01	295100.0	329050.0	681800.0	332200.0	184400.0	353900.0	490400.0	269350.0
2003-04-01	299250.0	336700.0	685500.0	333800.0	184100.0	359900.0	490200.0	271350.0
2003-05-01	305150.0	340300.0	686100.0	335900.0	183900.0	365700.0	489800.0	273400.0
2003-06-01	312750.0	343650.0	686700.0	338200.0	183700.0	371400.0	490300.0	276100.0
2003-07-01	321500.0	345900.0	687900.0	340600.0	183550.0	377600.0	492600.0	279050.0
2003-08-01	330550.0	349100.0	690300.0	343000.0	183350.0	384500.0	495700.0	282150.0
2003-09-01	339150.0	352500.0	694700.0	345800.0	183350.0	392600.0	500000.0	285300.0
2003-10-01	346850.0	354500.0	701300.0	349000.0	183450.0	399900.0	505200.0	288400.0
2003-11-01	353300.0	359400.0	710000.0	352600.0	183850.0	407700.0	509200.0	291350.0
2003-12-01	358600.0	365200.0	721100.0	356500.0	184750.0	415900.0	514100.0	294100.0
2004-01-01	363600.0	380700.0	734000.0	360400.0	186150.0	424500.0	519700.0	296700.0
2004-02-01	368800.0	382000.0	747800.0	364100.0	187700.0	434200.0	526100.0	299200.0
2004-03-01	374400.0	383200.0	762000.0	367600.0	189000.0	445600.0	533300.0	301700.0

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
2004-04-01	380700.0	385600.0	776700.0	370900.0	190200.0	465200.0	541700.0	304350.0
2004-05-01	387500.0	390700.0	791800.0	374000.0	191100.0	477500.0	550900.0	307350.0
2004-06-01	394700.0	396800.0	806600.0	376800.0	191750.0	488500.0	560600.0	310750.0
2004-07-01	402150.0	403600.0	820400.0	379600.0	192300.0	498900.0	570400.0	314400.0
2004-08-01	409700.0	411000.0	832600.0	384800.0	192500.0	508500.0	580000.0	318100.0
2004-09-01	418050.0	419200.0	843300.0	391100.0	192650.0	519700.0	589900.0	321650.0
2004-10-01	427250.0	428500.0	853000.0	396900.0	193000.0	529900.0	599600.0	325150.0
2004-11-01	436800.0	435700.0	861800.0	401600.0	193700.0	536100.0	610100.0	328400.0
2004-12-01	446650.0	440900.0	870100.0	406000.0	194500.0	542500.0	621500.0	331500.0
2005-01-01	456650.0	449300.0	877600.0	410600.0	195450.0	549400.0	633400.0	334650.0
2005-02-01	466050.0	455050.0	883800.0	415500.0	196300.0	556900.0	645500.0	337800.0
2005-03-01	472450.0	464750.0	889100.0	420800.0	197150.0	565100.0	657000.0	341100.0
2005-04-01	479000.0	470100.0	893900.0	425700.0	198000.0	574100.0	667000.0	344650.0
2005-05-01	485350.0	475550.0	898700.0	430300.0	198900.0	584000.0	676100.0	348100.0
2005-06-01	491100.0	481300.0	901300.0	434900.0	199800.0	593500.0	684700.0	351350.0
2005-07-01	494350.0	487450.0	901000.0	439500.0	200650.0	602100.0	692300.0	354400.0
2005-08-01	497200.0	495900.0	901100.0	443900.0	201250.0	610200.0	698400.0	357100.0
2005-09-01	499550.0	506700.0	903900.0	449200.0	201650.0	617900.0	702900.0	359700.0
2005-10-01	501400.0	517200.0	914700.0	455400.0	201900.0	624800.0	705900.0	362200.0
2005-11-01	502950.0	526050.0	923600.0	460900.0	202150.0	630000.0	707700.0	365500.0
2005-12-01	504550.0	533850.0	929200.0	466200.0	202500.0	633800.0	709000.0	368450.0

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
2006-01-01	506550.0	539650.0	932200.0	471500.0	202950.0	636700.0	708800.0	371850.0
2006-02-01	508850.0	544150.0	932100.0	477000.0	203650.0	639300.0	707600.0	375100.0
2006-03-01	510800.0	548950.0	924900.0	483000.0	204450.0	643800.0	706700.0	378100.0
2006-04-01	511950.0	553800.0	917200.0	489400.0	205250.0	648900.0	705500.0	380750.0
2006-05-01	512200.0	555900.0	909100.0	493800.0	206000.0	652700.0	705600.0	382350.0
2006-06-01	512100.0	558750.0	901100.0	497100.0	206600.0	654100.0	707400.0	382950.0
2006-07-01	511950.0	561750.0	893700.0	500000.0	207150.0	655300.0	707800.0	383150.0
2006-08-01	511800.0	564400.0	886700.0	502400.0	207900.0	653600.0	707900.0	383100.0
2006-09-01	511450.0	564800.0	880700.0	504500.0	208800.0	650100.0	708500.0	383050.0
2006-10-01	510900.0	563800.0	875900.0	506400.0	209750.0	648100.0	709400.0	383150.0
2006-11-01	510450.0	559800.0	874000.0	508000.0	210600.0	642200.0	710300.0	383150.0
2006-12-01	510600.0	554250.0	874900.0	509900.0	211150.0	637200.0	711300.0	383150.0
2007-01-01	511750.0	554950.0	878000.0	512800.0	211400.0	632500.0	712800.0	384100.0
2007-02-01	513500.0	552600.0	883200.0	516300.0	211500.0	631300.0	714300.0	384600.0
2007-03-01	515050.0	552400.0	893800.0	518900.0	211750.0	628800.0	714500.0	385150.0
2007-04-01	516100.0	555050.0	907200.0	521300.0	212300.0	624400.0	713600.0	386250.0
2007-05-01	516800.0	555200.0	920900.0	523100.0	212850.0	619800.0	713200.0	386800.0
2007-06-01	517250.0	554400.0	933700.0	524200.0	213400.0	614700.0	711700.0	386750.0
2007-07-01	517100.0	553100.0	944300.0	524600.0	214000.0	608900.0	708900.0	386850.0
2007-08-01	516650.0	549850.0	951200.0	524700.0	214500.0	601800.0	705200.0	387050.0
2007-09-01	515650.0	547100.0	954400.0	524500.0	214900.0	595300.0	700400.0	387500.0

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
2007-10-01	514150.0	544850.0	954800.0	523700.0	215100.0	587200.0	694400.0	388500.0
2007-11-01	511850.0	543650.0	952400.0	522300.0	214950.0	579300.0	687500.0	389100.0
2007-12-01	508850.0	544450.0	949200.0	520700.0	214550.0	571800.0	680300.0	389000.0
2008-01-01	505550.0	542650.0	946100.0	518600.0	213500.0	564900.0	673900.0	387850.0
2008-02-01	502200.0	538150.0	943200.0	515300.0	212150.0	556800.0	668400.0	385050.0
2008-03-01	498800.0	533900.0	940700.0	511100.0	210950.0	546700.0	660700.0	381150.0
2008-04-01	495350.0	533100.0	937900.0	506900.0	210000.0	534900.0	650100.0	375850.0
2008-05-01	492450.0	527100.0	933100.0	503000.0	209400.0	524900.0	639600.0	370150.0
2008-06-01	490150.0	521700.0	926800.0	499200.0	209300.0	513400.0	631800.0	365750.0
2008-07-01	487950.0	515750.0	919800.0	493500.0	209600.0	505000.0	624200.0	362150.0
2008-08-01	485300.0	511200.0	909200.0	488000.0	210000.0	499800.0	616200.0	357900.0
2008-09-01	482250.0	506750.0	900200.0	483300.0	210600.0	493100.0	608600.0	354100.0
2008-10-01	478950.0	502500.0	890200.0	479100.0	211350.0	487400.0	601400.0	351250.0
2008-11-01	475850.0	500050.0	879300.0	475600.0	212100.0	483100.0	592100.0	348750.0
2008-12-01	473050.0	498150.0	870300.0	472700.0	213000.0	479400.0	586400.0	345950.0
2009-01-01	470700.0	494900.0	864700.0	470000.0	213950.0	476400.0	580600.0	344000.0
2009-02-01	468450.0	491250.0	857700.0	465600.0	214750.0	473000.0	573300.0	341150.0
2009-03-01	465650.0	487350.0	847900.0	459600.0	215200.0	469000.0	564400.0	337500.0
2009-04-01	462150.0	483450.0	835700.0	452500.0	215450.0	464300.0	553900.0	333550.0
2009-05-01	458500.0	479500.0	822900.0	444800.0	215450.0	456900.0	543200.0	329000.0
2009-06-01	455350.0	475300.0	811300.0	437800.0	215350.0	451100.0	534100.0	323750.0

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
2009-07-01	453250.0	473350.0	802200.0	433200.0	215350.0	448400.0	527300.0	319000.0
2009-08-01	452250.0	473250.0	796300.0	431500.0	215500.0	447500.0	523100.0	314450.0
2009-09-01	451750.0	472900.0	796300.0	430900.0	216000.0	448000.0	521100.0	309900.0
2009-10-01	451600.0	473150.0	797600.0	429200.0	216950.0	449800.0	520700.0	305550.0
2009-11-01	451950.0	474500.0	800100.0	428100.0	218150.0	451800.0	521700.0	303900.0
2009-12-01	453500.0	477350.0	808300.0	428400.0	219100.0	455700.0	524000.0	303950.0
2010-01-01	455250.0	478650.0	818800.0	429500.0	219950.0	458800.0	528500.0	303100.0
2010-02-01	455850.0	478000.0	826000.0	429400.0	220950.0	458600.0	530400.0	303200.0
2010-03-01	456850.0	478850.0	823400.0	426900.0	220750.0	456500.0	530100.0	303300.0
2010-04-01	458400.0	474850.0	815100.0	426100.0	219550.0	459700.0	532800.0	301700.0
2010-05-01	457950.0	476550.0	821100.0	427600.0	219450.0	459900.0	544700.0	299050.0
2010-06-01	454850.0	475450.0	832000.0	429200.0	219350.0	459000.0	545500.0	299200.0
2010-07-01	450350.0	475200.0	836600.0	428000.0	218200.0	457500.0	544000.0	298350.0
2010-08-01	445100.0	473450.0	843900.0	425500.0	217500.0	454200.0	547900.0	295650.0
2010-09-01	441250.0	471300.0	847700.0	423100.0	217400.0	448000.0	550500.0	292400.0
2010-10-01	439200.0	463050.0	849900.0	420300.0	215700.0	442600.0	551100.0	288450.0
2010-11-01	438150.0	456800.0	854600.0	415800.0	214300.0	436600.0	548200.0	285700.0
2010-12-01	437500.0	456050.0	851300.0	409100.0	213650.0	431700.0	542300.0	283800.0
2011-01-01	436950.0	454550.0	843200.0	403000.0	213100.0	427000.0	536500.0	282750.0
2011-02-01	435400.0	454850.0	835700.0	400000.0	212750.0	425700.0	532700.0	280900.0
2011-03-01	432500.0	463950.0	830500.0	399200.0	213250.0	425400.0	530900.0	278350.0

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
2011-04-01	430800.0	475750.0	824300.0	396100.0	213600.0	424700.0	528000.0	275000.0
2011-05-01	431650.0	473400.0	815600.0	393400.0	212900.0	422700.0	525300.0	271950.0
2011-06-01	434350.0	468350.0	804000.0	392800.0	211750.0	422800.0	523600.0	270200.0
2011-07-01	437500.0	464000.0	795700.0	394200.0	210450.0	421000.0	521800.0	269850.0
2011-08-01	441150.0	462650.0	794500.0	394600.0	209650.0	417300.0	519700.0	270050.0
2011-09-01	445200.0	468650.0	802200.0	393600.0	209750.0	414300.0	518300.0	269650.0
2011-10-01	449350.0	474350.0	813100.0	392900.0	210650.0	412800.0	518500.0	267900.0
2011-11-01	452500.0	474300.0	820600.0	394800.0	211550.0	409200.0	519300.0	265400.0
2011-12-01	454150.0	468600.0	824200.0	397800.0	212500.0	405200.0	516600.0	264650.0
2012-01-01	454050.0	452400.0	825800.0	398700.0	213350.0	402200.0	513800.0	263150.0
2012-02-01	453700.0	449000.0	830800.0	399000.0	214250.0	403800.0	518700.0	259950.0
2012-03-01	455350.0	448300.0	838500.0	401000.0	215150.0	404800.0	526200.0	258200.0
2012-04-01	458300.0	447300.0	848100.0	405800.0	216150.0	408700.0	530700.0	259550.0
2012-05-01	461150.0	447000.0	858400.0	410700.0	216300.0	412700.0	535400.0	260750.0
2012-06-01	463750.0	448800.0	866000.0	416000.0	216200.0	416100.0	543700.0	260450.0
2012-07-01	466700.0	450900.0	873500.0	420100.0	217100.0	421400.0	553700.0	258950.0
2012-08-01	470150.0	450700.0	889200.0	422700.0	218400.0	424900.0	564400.0	258300.0
2012-09-01	473700.0	448400.0	904500.0	426000.0	219200.0	428700.0	572000.0	259200.0
2012-10-01	477150.0	451300.0	923800.0	431100.0	219750.0	431300.0	578700.0	261150.0
2012-11-01	479400.0	458900.0	941000.0	435700.0	220400.0	434000.0	586000.0	263200.0
2012-12-01	480300.0	465100.0	947100.0	439300.0	220700.0	437600.0	593300.0	264500.0

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
2013-01-01	482550.0	468800.0	954200.0	444200.0	220950.0	442600.0	599400.0	265450.0
2013-02-01	487250.0	471700.0	966800.0	450100.0	221200.0	448700.0	605800.0	267050.0
2013-03-01	492350.0	475400.0	987300.0	456600.0	221250.0	458100.0	611600.0	270000.0
2013-04-01	497250.0	480600.0	1013100.0	463100.0	221400.0	468200.0	620900.0	273700.0
2013-05-01	504300.0	483500.0	1032100.0	469500.0	222350.0	480900.0	636500.0	278050.0
2013-06-01	513000.0	491500.0	1042700.0	473200.0	223700.0	495300.0	653600.0	283900.0
2013-07-01	521650.0	493000.0	1050800.0	475300.0	224850.0	487100.0	665900.0	295000.0
2013-08-01	529300.0	492000.0	1064600.0	476500.0	225900.0	490500.0	673300.0	301500.0
2013-09-01	536100.0	498500.0	1081100.0	476400.0	226450.0	495200.0	679000.0	304700.0
2013-10-01	540750.0	504000.0	1094500.0	474400.0	226700.0	500700.0	682700.0	305400.0
2013-11-01	546700.0	509100.0	1107400.0	474500.0	227100.0	503700.0	685400.0	305100.0
2013-12-01	554300.0	515400.0	1125700.0	475800.0	228250.0	506000.0	689400.0	304300.0
2014-01-01	564900.0	522050.0	1148100.0	478500.0	215400.0	509100.0	695400.0	304800.0
2014-02-01	574700.0	518600.0	1170800.0	482100.0	215500.0	514000.0	700000.0	306800.0
2014-03-01	584650.0	521750.0	1187600.0	485100.0	216300.0	520500.0	703100.0	308700.0
2014-04-01	593750.0	523950.0	1202200.0	487300.0	217900.0	527700.0	709800.0	310200.0
2014-05-01	600900.0	524700.0	1221900.0	490900.0	219500.0	534500.0	715300.0	312100.0
2014-06-01	604600.0	526600.0	1240800.0	497000.0	221200.0	541900.0	716200.0	313800.0
2014-07-01	606400.0	531400.0	1258300.0	503600.0	223800.0	549000.0	720500.0	316400.0
2014-08-01	608850.0	535900.0	1270500.0	510200.0	226400.0	554000.0	728000.0	318400.0
2014-09-01	611050.0	537400.0	1277100.0	517000.0	230000.0	557700.0	735700.0	318700.0

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
2014-10-01	613400.0	538900.0	1287200.0	522700.0	234500.0	560100.0	743300.0	317900.0
2014-11-01	615750.0	544800.0	1300000.0	527100.0	238100.0	561000.0	751400.0	317300.0
2014-12-01	617800.0	550150.0	1304900.0	533200.0	239900.0	561300.0	759900.0	317400.0
2015-01-01	621050.0	552450.0	1309100.0	539700.0	241000.0	563800.0	768400.0	316800.0
2015-02-01	627950.0	553450.0	1318900.0	544900.0	242800.0	570700.0	777100.0	316000.0
2015-03-01	635500.0	554750.0	1334100.0	550000.0	245400.0	574600.0	784300.0	316900.0
2015-04-01	641900.0	557750.0	1351300.0	556200.0	248000.0	578600.0	792100.0	318300.0
2015-05-01	649250.0	564500.0	1369200.0	561600.0	250400.0	581800.0	803800.0	318000.0
2015-06-01	658100.0	573600.0	1389900.0	566200.0	253200.0	583000.0	816800.0	319000.0
2015-07-01	666500.0	579250.0	1414000.0	572000.0	255800.0	584800.0	825700.0	320700.0
2015-08-01	675100.0	584050.0	1435700.0	578300.0	257600.0	592700.0	832700.0	320700.0
2015-09-01	685100.0	586250.0	1448400.0	585200.0	258900.0	601600.0	838300.0	320700.0
2015-10-01	693550.0	587200.0	1452300.0	595600.0	259800.0	612700.0	842600.0	322100.0
2015-11-01	698600.0	589250.0	1454500.0	607700.0	261400.0	621300.0	846100.0	323600.0
2015-12-01	701700.0	592750.0	1457400.0	618700.0	263700.0	628400.0	851200.0	326600.0
2016-01-01	703700.0	598150.0	1465000.0	628600.0	266000.0	635600.0	859200.0	329900.0
2016-02-01	705000.0	602750.0	1474400.0	638600.0	267400.0	637200.0	867500.0	332300.0
2016-03-01	706200.0	604850.0	1480200.0	648400.0	268500.0	639300.0	872600.0	333700.0
2016-04-01	708550.0	609400.0	1477700.0	656800.0	270100.0	641000.0	876700.0	334600.0
2016-05-01	710550.0	615900.0	1475800.0	665000.0	272400.0	643200.0	882900.0	336600.0
2016-06-01	711300.0	622500.0	1472500.0	673800.0	275500.0	646500.0	891100.0	339300.0

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
2016-07-01	711100.0	630100.0	1463300.0	682500.0	280000.0	651400.0	898800.0	340900.0
2016-08-01	710350.0	637550.0	1453400.0	689100.0	285800.0	657100.0	903500.0	342300.0
2016-09-01	708650.0	644150.0	1457100.0	694800.0	292600.0	664200.0	906200.0	344000.0
2016-10-01	708350.0	650350.0	1470400.0	701400.0	299700.0	670600.0	909000.0	346400.0
2016-11-01	710900.0	657450.0	1487400.0	709100.0	305700.0	664100.0	910700.0	348800.0
2016-12-01	717300.0	660300.0	1502200.0	715100.0	308800.0	661600.0	915300.0	349900.0
2017-01-01	724600.0	663800.0	1514300.0	721600.0	309400.0	666400.0	923000.0	350300.0
2017-02-01	729450.0	666300.0	1525000.0	730800.0	309900.0	667000.0	932100.0	350700.0
2017-03-01	732600.0	669650.0	1527200.0	740500.0	311600.0	661900.0	939800.0	351300.0
2017-04-01	736750.0	674150.0	1534900.0	752600.0	314600.0	668400.0	942700.0	352800.0
2017-05-01	741050.0	680900.0	1552300.0	767600.0	316900.0	677200.0	945700.0	354500.0
2017-06-01	744600.0	687750.0	1570600.0	781000.0	317800.0	686000.0	950700.0	357200.0
2017-07-01	748900.0	694800.0	1600000.0	789600.0	319600.0	694500.0	961000.0	355500.0
2017-08-01	755250.0	703500.0	1625600.0	794300.0	322300.0	702100.0	976100.0	356200.0
2017-09-01	761050.0	707000.0	1648500.0	798200.0	325000.0	708000.0	998100.0	357100.0
2017-10-01	764800.0	705600.0	1675200.0	804900.0	325900.0	714900.0	1024100.0	358000.0
2017-11-01	768450.0	713000.0	1690000.0	814400.0	326200.0	723400.0	1051900.0	358500.0
2017-12-01	771150.0	727350.0	1704500.0	827100.0	326700.0	730500.0	1077100.0	355200.0
2018-01-01	772550.0	735300.0	1719100.0	840200.0	328000.0	735600.0	1096900.0	353900.0
2018-02-01	776650.0	740250.0	1737600.0	847000.0	330800.0	739900.0	1119000.0	356400.0
2018-03-01	783050.0	743050.0	1757900.0	849600.0	333900.0	743600.0	1149300.0	356900.0

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
2018-04-01	785750.0	743950.0	1772500.0	850400.0	335100.0	762500.0	1170100.0	356200.0

In [22]:
```
# data frame is clean, no nulls
melted_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 265 entries, 1996-04-01 to 2018-04-01
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Washington, DC     265 non-null    float64
 1   New York, NY       265 non-null    float64
 2   San Francisco, CA  265 non-null    float64
 3   Seattle, WA        265 non-null    float64
 4   Dallas, TX         265 non-null    float64
 5   Los Angeles, CA    265 non-null    float64
 6   San Jose, CA       265 non-null    float64
 7   Chicago, IL        265 non-null    float64
 8   Baltimore, MD      265 non-null    float64
 9   Boston, MA         265 non-null    float64
dtypes: float64(10)
memory usage: 32.8 KB
```

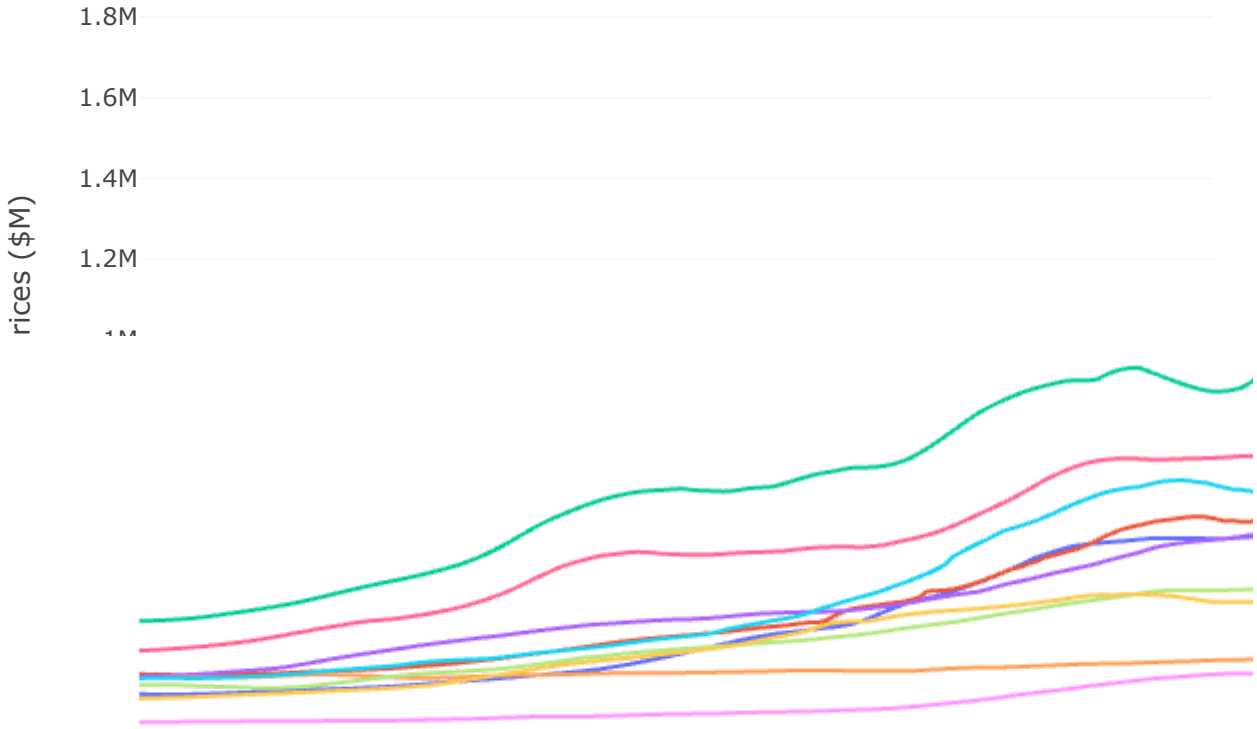
In [23]:
```
# created a yearly data frame for use in our baseline model
resampled_year = melted_df.resample('A').median()
```

EDA Time series data

While the graph with annual data is a little more smooth than the graph with monthly data (as expected), the graphs still look similiar enough to continue to send accross the same message. In fact, the less noisy annual data should slightly increase the root mean squared error for the baseline models.

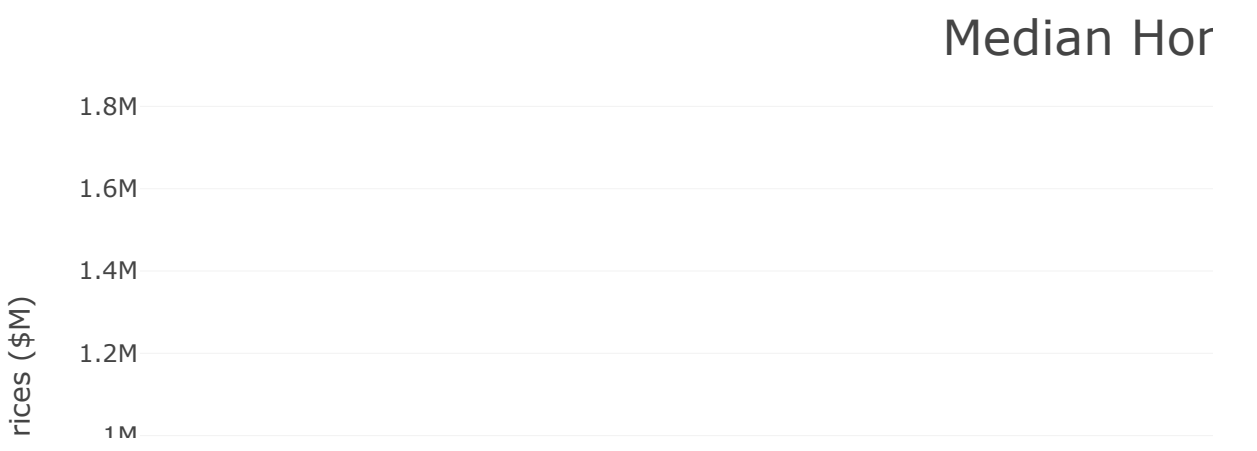
Plot of monthly data from 1996 - 2018

In [24]:
```
# Plot median house price time series for each City:
fig = px.line(melted_df, labels={"variable": "City", "value": "Median Home Price

fig.update_layout({"plot_bgcolor": "rgba(0, 0, 0, 0)",
                   "paper_bgcolor": "rgba(0, 0, 0, 0)"},
                  title_text = 'Median Home Prices (1996 - 2018)',
                  title_font_size = 24,
                  title_xref = 'container',
                  title_y = 0.95,
                  title_x = 0.5,
                  hovermode = 'closest',
                  template = 'xgridoff')
fig.show()
```



Plot of Yearly data from 1996-2018

In [25]:
```
# Plot median house price time series for each City:
fig = px.line(resampled_year, labels={"variable": "City", "value": "Median Home

fig.update_layout(title_text = 'Median Home Prices (1996 - 2018)',
                  title_font_size = 24,
                  title_xref = 'container',
                  title_y = 0.95,
                  title_x = 0.5,
                  hovermode = 'closest',
                  template = 'xgridoff')
fig.show()
```



Return of investments by city

In [26]:
```
# Calculate return for each City:
df_return = melted_df.pct_change()
df_return

# drop the first row of the df_return dataframe:
df_return.dropna(axis=0, inplace=True)

df_return.head()
```

Out[26]:

time	Washington, DC	New York, NY	San Francisco, CA	Seattle, WA	Dallas, TX	Los Angeles, CA	San Jose, CA	Chicago, IL
1996-05-01	-0.001976	-0.001705	0.002281	0.002345	0.005443	0.004834	0.002131	-0.000334
1996-06-01	-0.001980	0.000854	0.002601	0.003509	0.003609	0.003608	0.003403	-0.001670
1996-07-01	-0.001587	-0.003697	0.002918	0.003497	0.001498	-0.002397	0.005935	-0.002342
1996-08-01	-0.000397	-0.004282	0.003880	0.004646	0.000898	-0.001802	0.006321	-0.000335
1996-09-01	0.001193	-0.002294	0.004831	0.004624	0.000598	-0.001805	0.006700	-0.002684

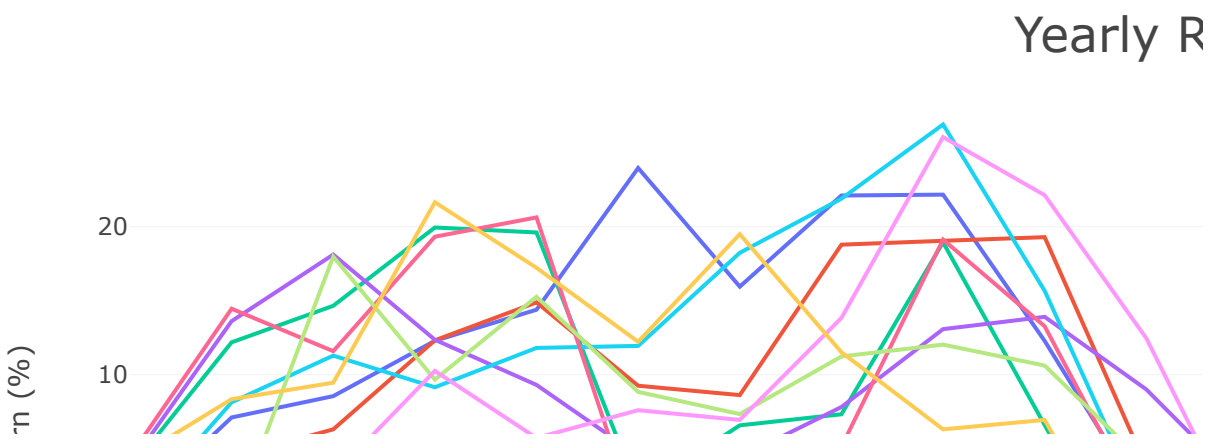
We see that most of our cities have very high average annual returns. The S&P 500 averaged about 10% annual returns over these years, and it would be easy to match those numbers for several of these cities given smart investments were made. It could also be wise to invest in some of these slower growth cities as they are poised for growth as tech takes hold in them.

In [27]:
```
#Create yearly return dataframe

#Pull out Year
df_yearly_return = df_return
df_yearly_return['date'] = list(df_return.index)
df_yearly_return['year'] = df_yearly_return.date.dt.year
```

```
#Group by year, perform aggregation
df_yearly_return = df_yearly_return.groupby('year').sum()*100
df_yearly_return.mean()
```

Out[27]:
```
Washington, DC       8.004704
New York, NY         6.328264
San Francisco, CA    7.692601
Seattle, WA          7.037208
Dallas, TX           3.095600
Los Angeles, CA      6.721343
San Jose, CA         7.059666
Chicago, IL          3.810362
Baltimore, MD        4.760559
Boston, MA           7.057949
dtype: float64
```

In [28]:
```
# Plot yearly return for the last 2 decades for each City:
fig = px.line(df_yearly_return, labels={"variable": "City", "value": "Yearly Ret

fig.update_layout(title_text = 'Yearly Return (1996 - 2018)',
                  title_font_size = 24,
                  title_xref = 'container',
                  title_y = 0.95,
                  title_x = 0.5,
                  hovermode = 'closest',
                  template = 'xgridoff')
fig.show()
```



In [29]:
```
# Construct new dataframe for EDA purposes:
#mean monthly return
df_cum = pd.DataFrame(data=df_return.mean())
df_cum.rename(columns = {0:'MonthlyReturnMean'}, inplace = True)

#cumulative return
cumsum = []
for i in df_cum.index:
    cumsum.append(df_return[i].cumsum()[-1])
df_cum['CumulativeReturn'] = cumsum

# cumulative return %
df_cum['CumulativeReturn(%)'] = df_cum['CumulativeReturn']*100
df_cum
# Average Yearly Return (%)
df_cum['AverageYearlyReturn(%)'] = df_cum['CumulativeReturn']*100/22 # We have 9

# reset the index
df_cum = df_cum.reset_index()
# name the index - city
df_cum.rename(columns = {'index':'City'}, inplace = True)

df_cum
```

```
<ipython-input-29-ecfcba425b17>:3: FutureWarning:

DataFrame.mean and DataFrame.median with numeric_only=None will include datetime
64 and datetime64tz columns in a future version.
```

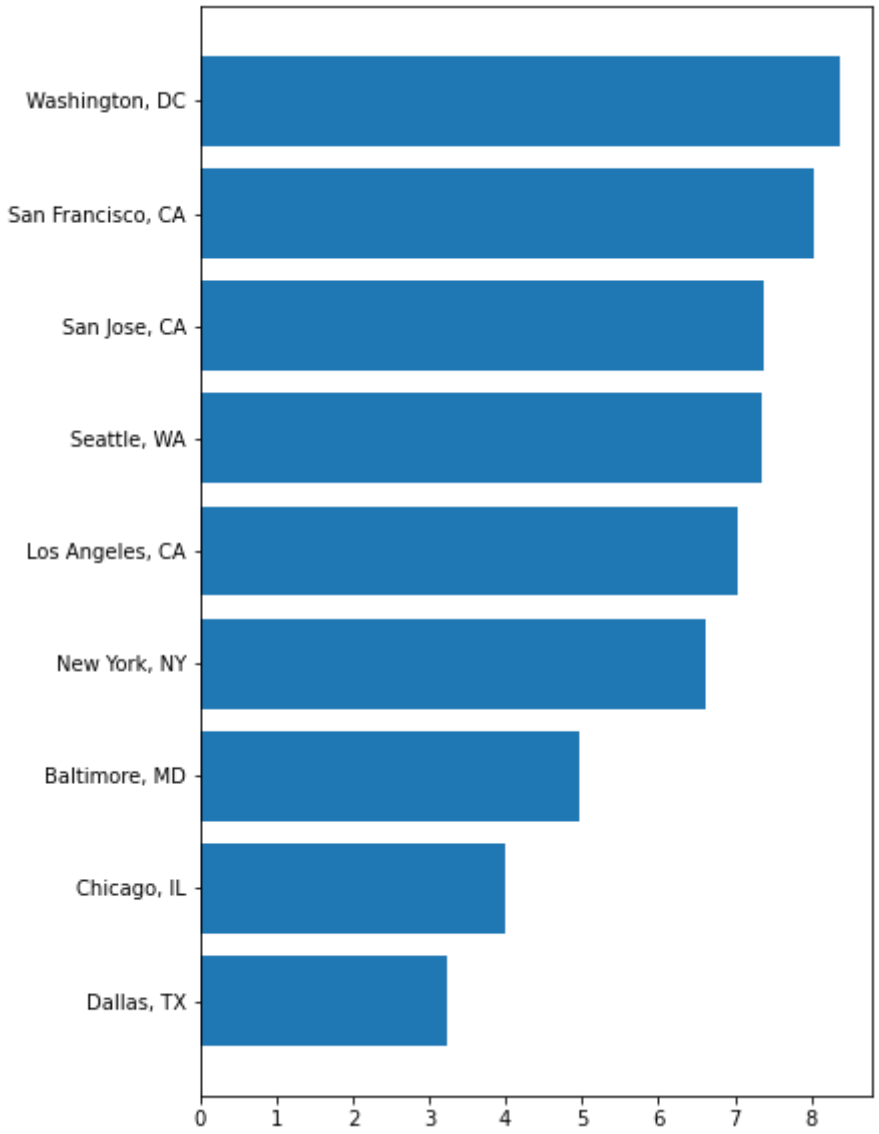
Out[29]:

	City	MonthlyReturnMean	CumulativeReturn	CumulativeReturn(%)	AverageYearlyRetur
0	Washington, DC	0.006974	1.841082	1.841082e+02	8.368554
1	New York, NY	0.005513	1.455501	1.455501e+02	6.615913
2	San Francisco, CA	0.006702	1.769298	1.769298e+02	8.042264
3	Seattle, WA	0.006131	1.618558	1.618558e+02	7.357081
4	Dallas, TX	0.002697	0.711988	7.119880e+01	3.236309
5	Los Angeles, CA	0.005856	1.545909	1.545909e+02	7.026859
6	San Jose, CA	0.006150	1.623723	1.623723e+02	7.380560
7	Chicago, IL	0.003320	0.876383	8.763833e+01	3.983560
8	Baltimore, MD	0.004147	1.094929	1.094929e+02	4.976948
9	Boston, MA	0.006149	1.623328	1.623328e+02	7.378764
10	year	2006.833333	529804.000000	5.298040e+07	2.408200

In [30]:
```
fig, ax = plt.subplots(figsize=(6,10))
```

```
cum_sorted = df_cum[:-2].sort_values(by=['AverageYearlyReturn(%)'], ascending =
plt.barh(y=cum_sorted['City'], width=cum_sorted['AverageYearlyReturn(%)'])
```

Out[30]: <BarContainer object of 9 artists>



Model Building

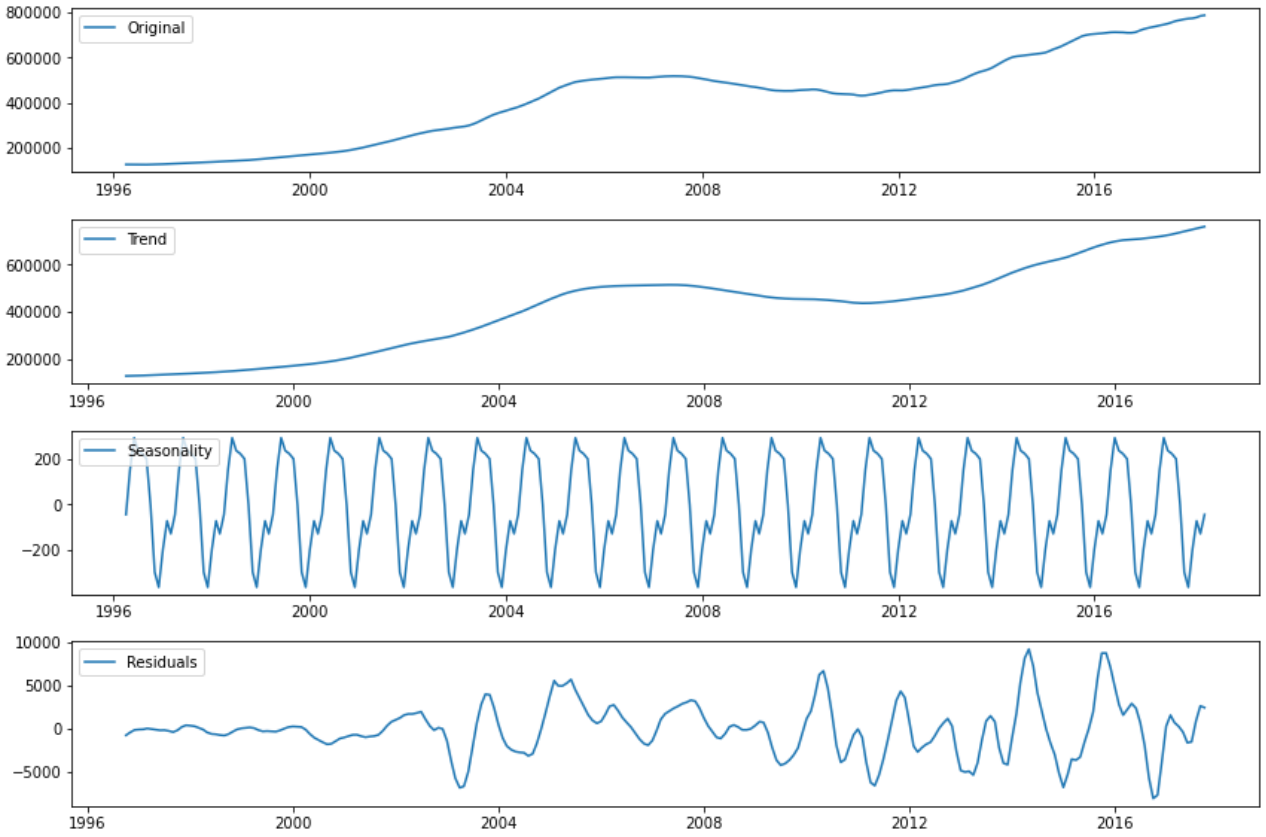
Time series data decomposition

Here we wanted to see how different constituents of the time series breakdown after decomposition. As it seemed from the graphs above, the vast majority of the variance in the values was captured in the overall trend. It seems that seasonality really effects the prices of homes, as the scale of the seasonal decompostion was on the scale of hundreds compared to the scale of house prices being of the magnitude of hundreds of thousands. Due to the lack of seasonality, an ARIMA model should be appropriate. There was a little noise left after the decomposition, but it looked a little off. However, the residuals were considered stationary according to the Dickey Fuller test.

In [31]:
```
#Decomposing
decomposition = seasonal_decompose(melted_df['Washington, DC'])
```

```
#General trend. (i.e. overtime,how does the house market change.)
trend = decomposition.trend

#Seasonal trend
seasonal = decomposition.seasonal

#This will be the leftover noise in the model.
residual = decomposition.resid
```

In [32]:
```
# Plot gathered statistics
plt.figure(figsize=(12, 8))
plt.subplot(411)
plt.plot(melted_df['Washington, DC'], label='Original')
plt.legend(loc='upper left')
plt.subplot(412)
plt.plot(trend, label='Trend')
plt.legend(loc='upper left')
plt.subplot(413)
plt.plot(seasonal,label='Seasonality')
plt.legend(loc='upper left')
plt.subplot(414)
plt.plot(residual, label='Residuals')
plt.legend(loc='upper left')
plt.tight_layout()
```



In [33]:
```
# Drop NaN values from residuals.
house_ts_decompose = residual
house_ts_decompose.dropna(inplace = True)
```

In [34]:
```
#Obtained this function from lecture 58 notebook
```

```
def display_df(dftest):
    '''
    Display the output from a Dickey-Fuller test in a more readable format
    '''


    dfoutput = pd.Series(
                dftest[0:4],
                index=['Test Statistic','p-value','#Lags Used','Number of Observ

    for key,value in dftest[4].items():
        dfoutput['Critical Value (%s)'%key] = value

    display(dfoutput)
```

In [35]:
```
#After we take out the trend and the seasonality, we look at the residuals with
dftest = adfuller(house_ts_decompose)

#Print out our results.
display_df(dftest)
```

```
Test Statistic                  -4.719574
p-value                          0.000077
#Lags Used                      10.000000
Number of Observations Used    242.000000
Critical Value (1%)             -3.457664
Critical Value (5%)             -2.873559
Critical Value (10%)            -2.573175
dtype: float64
```

Baseline Model

Here we created a class that could get all of the relevant information from our cities. The plot_shift method will plot all of the one year shifts on a single axes. The shift df method will return a df that with three shifted periods for 1, 2, and 3 years, as well give root mean squared errors for the three different baselines. We plan on forecasting 3 years into the future, but knowing how good numbers to beat for shorter time periods is also useful. After running all the baseline models, our average root mean squared error was about $118,000.

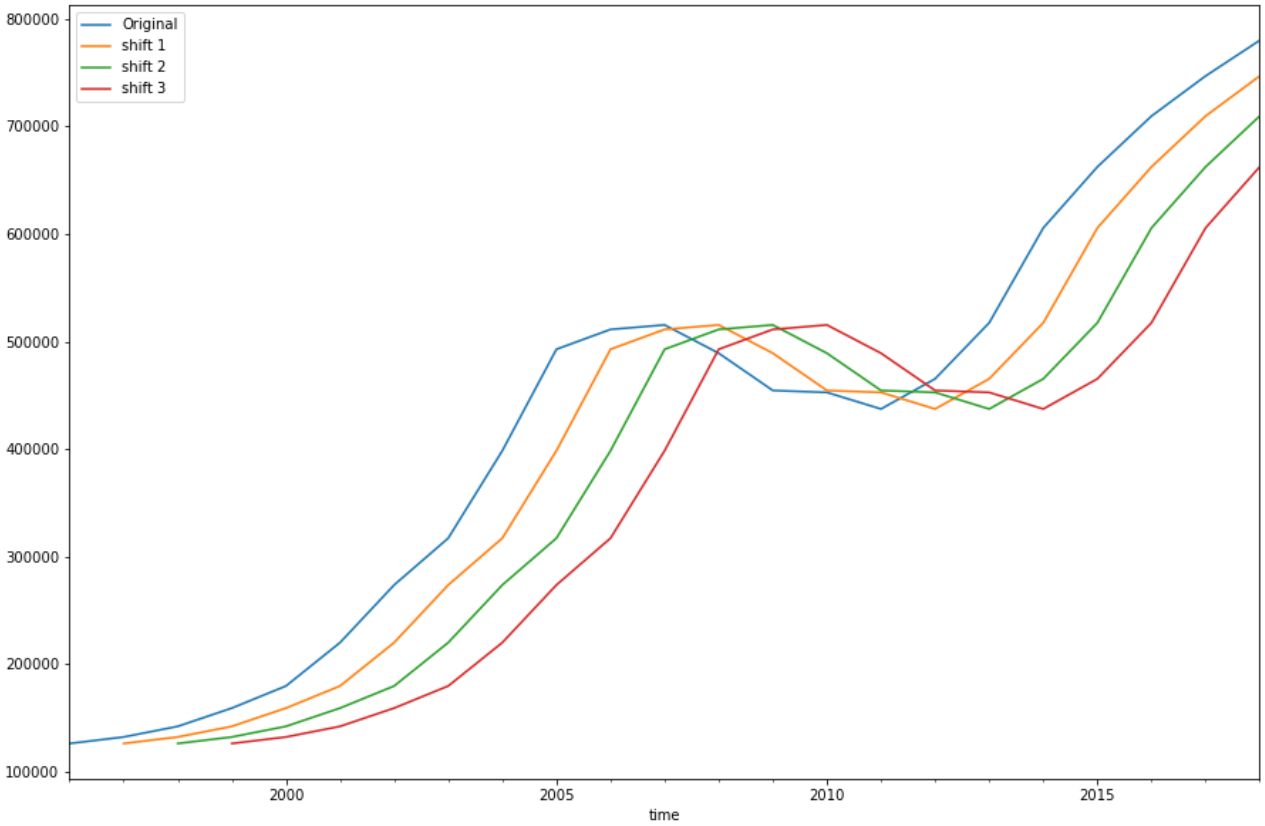
In [36]:
```
class baseline_mod:

    def __init__(self,city='Washington, DC'):
        self.city = city

    def plot_shift(self,df):
        ax = df[self.city].plot(figsize=(15,10))
        df[self.city].shift(1).plot()
        df[self.city].shift(2).plot()
        df[self.city].shift(3).plot()
        ax.legend(['Original', 'shift 1', 'shift 2', 'shift 3'])
        plt.show()


    def shift_df(self, df):
        self.shifted_df = pd.DataFrame(np.hstack((df[self.city].values.reshape(-
                                                  df[self.city].shift().values.r
                                                  df[self.city].shift(periods=2)
                                                  df[self.city].shift(periods=3)
```

```
                                        columns=['orig', 'shifted_one_period', 's
                                        index=df.index)
        self.rmse_shift1_ = mean_squared_error(self.shifted_df['orig'][1:], self
        self.rmse_shift2_ = mean_squared_error(self.shifted_df['orig'][2:], self
        self.rmse_shift3_ = mean_squared_error(self.shifted_df['orig'][3:], self


        return self.shifted_df
```

Washington D.C.

In [37]:
```
# graphing the baseline model
def baseline_graph(city):
    baseline_year = baseline_mod(city=city)
    baseline_year.plot_shift(resampled_year)
```

In [38]:
```
baseline_graph('Washington, DC')
```



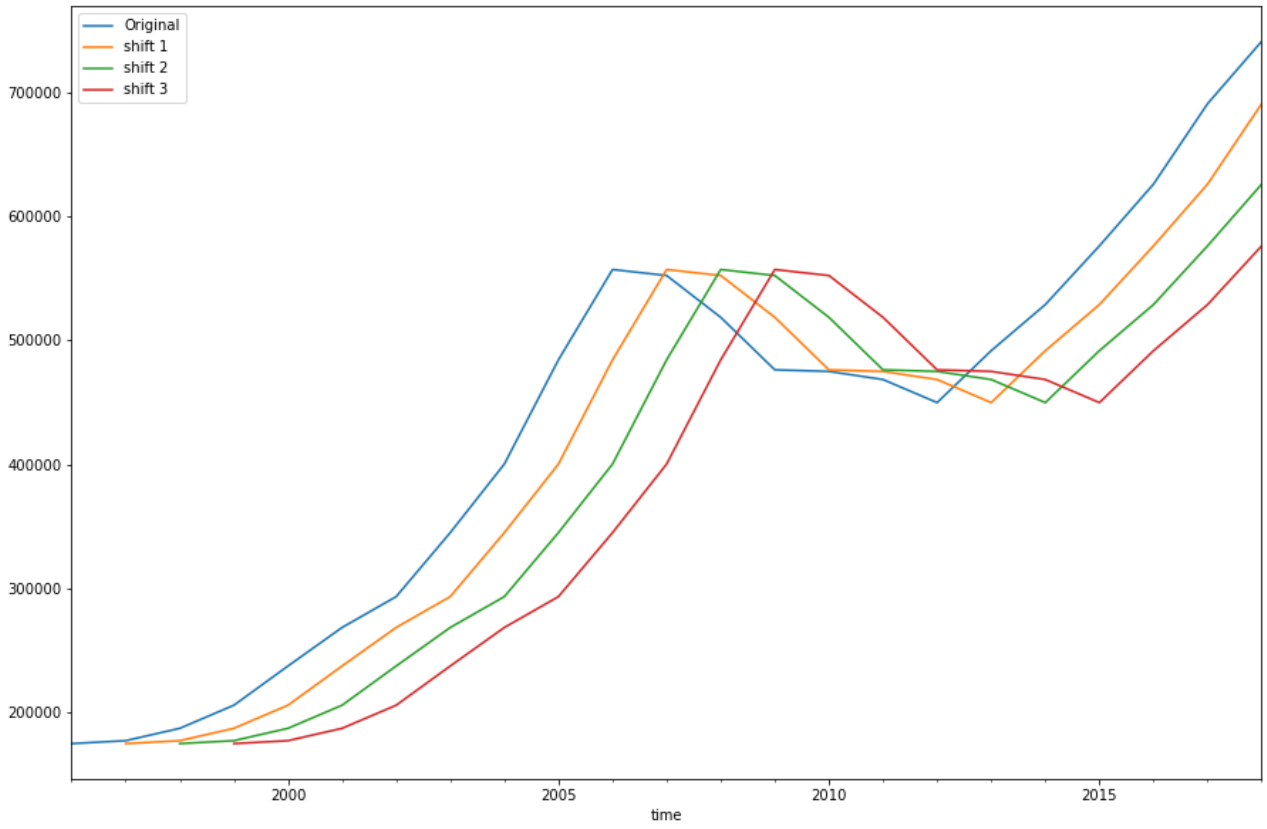
In [39]:
```
# make a function to quickly get see the rmse
def see_rmse(city):
    baseline_year = baseline_mod(city=city)
    baseline_year.shift_df(resampled_year)
    print(f"RMSE shifted 1 year:{baseline_year.rmse_shift1_}")
    print(f"RMSE shifted 2 years:{baseline_year.rmse_shift2_}")
    print(f"RMSE shifted 3 years:{baseline_year.rmse_shift3_}")
# and get the rmse
def get_rmse(city):
    baseline_year = baseline_mod(city=city)
    baseline_year.shift_df(resampled_year)
    return baseline_year.rmse_shift3_
```

In [40]:
```
baseline_rmse = []
washington_rmse = get_rmse('Washington, DC')
baseline_rmse.append(washington_rmse)
see_rmse('Washington, DC')
```

```
RMSE shifted 1 year:44942.247030848586
RMSE shifted 2 years:87278.32651788132
RMSE shifted 3 years:127389.16898622112
```

New York

In [41]:
```
baseline_graph('New York, NY')
```

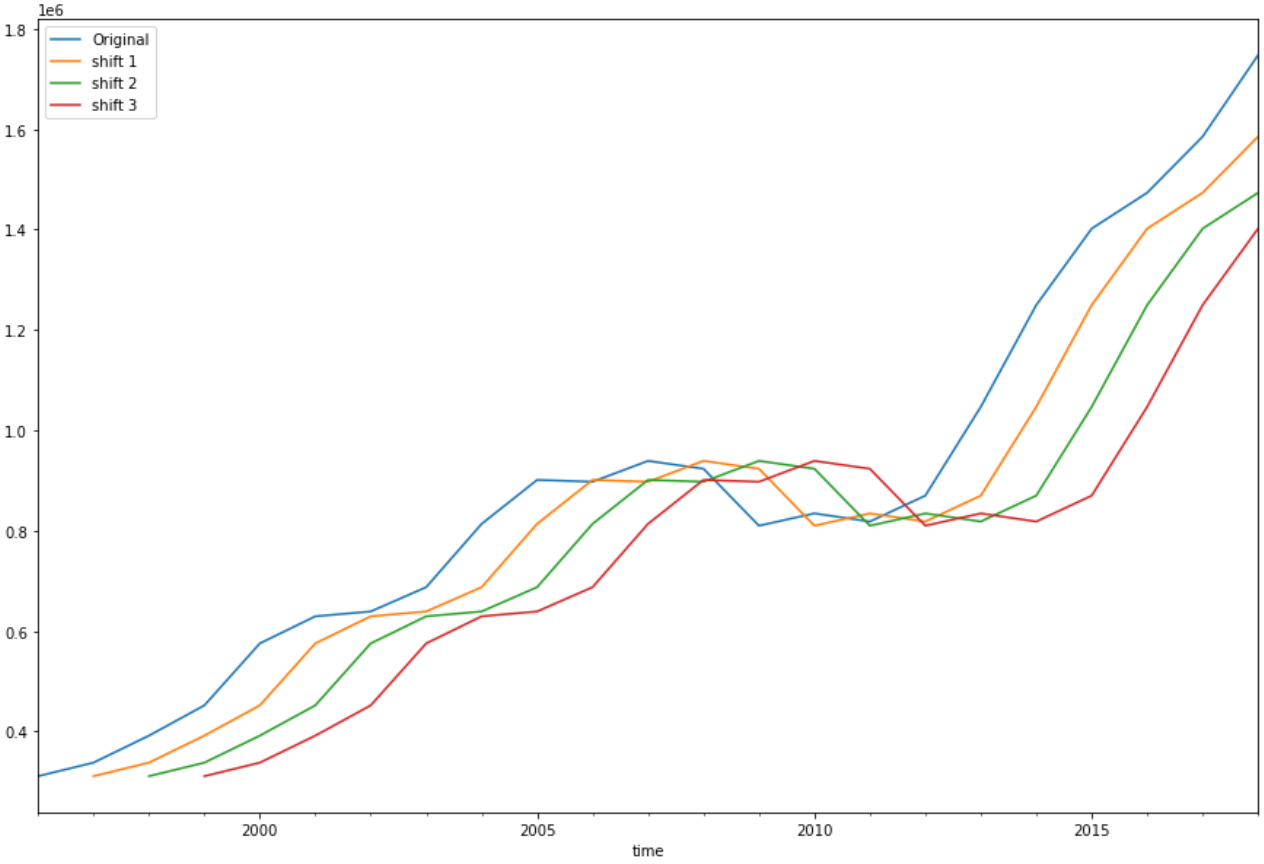


In [42]:
```
new_york_rmse = get_rmse('New York, NY')
baseline_rmse.append(new_york_rmse)
see_rmse('New York, NY')
```

```
RMSE shifted 1 year:42204.974663700916
RMSE shifted 2 years:80647.0468195475
RMSE shifted 3 years:114366.02205200634
```

San Francisco

In [43]:
```
baseline_graph('San Francisco, CA')
```

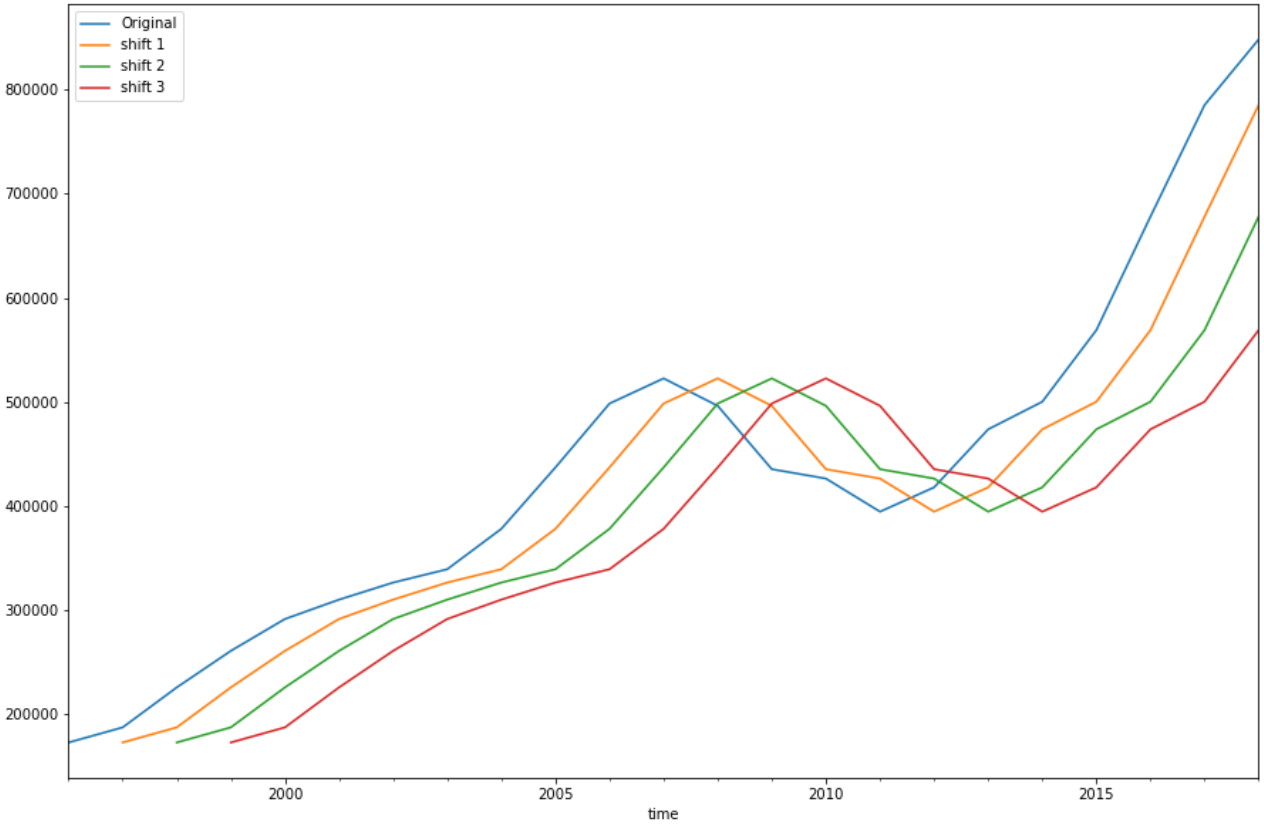


In [44]:
```
san_francisco_rmse = get_rmse('San Francisco, CA')
baseline_rmse.append(san_francisco_rmse)
see_rmse('San Francisco, CA')
```

```
RMSE shifted 1 year:97600.79347199814
RMSE shifted 2 years:180645.4877983526
RMSE shifted 3 years:255531.38310195872
```

Seattle

In [45]:
```
baseline_graph('Seattle, WA')
```

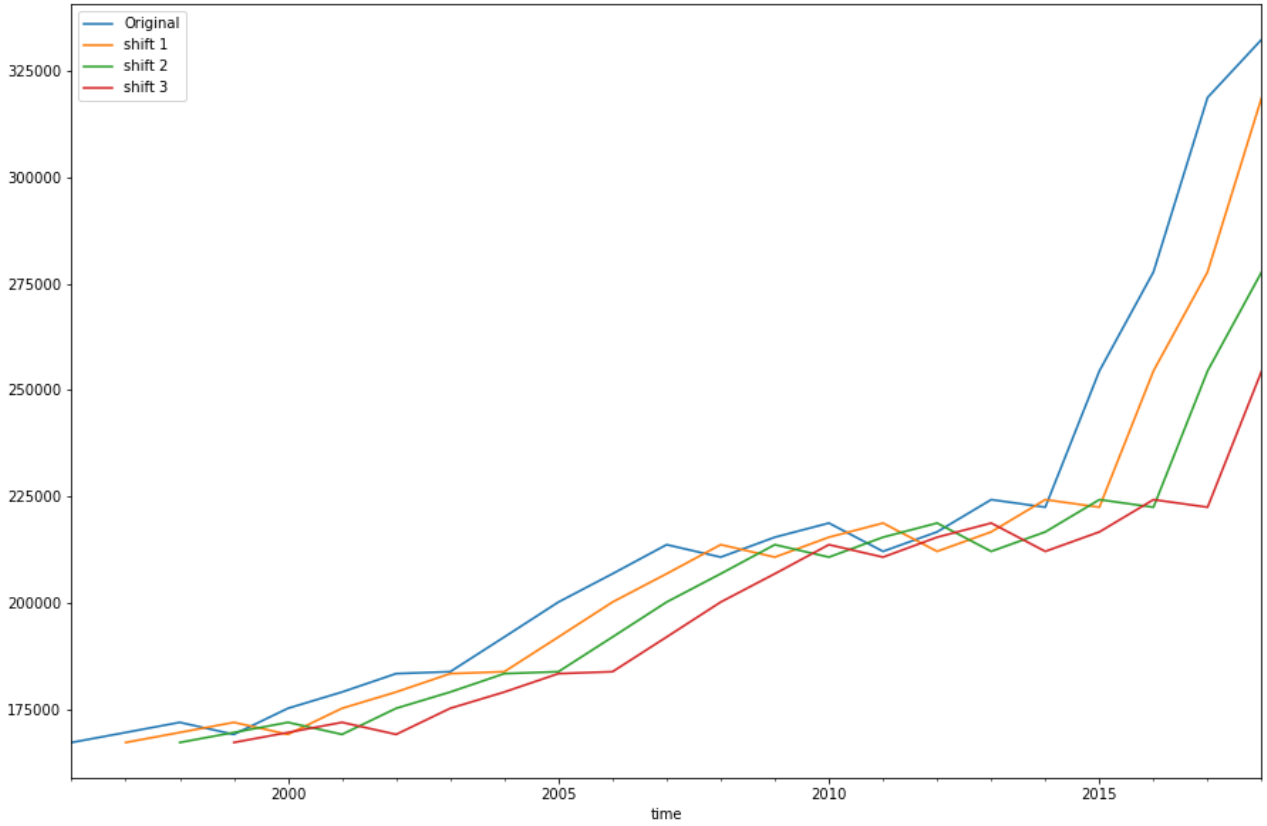


In [46]:
```
seattle_rmse = get_rmse('Seattle, WA')
baseline_rmse.append(seattle_rmse)
see_rmse('Seattle, WA')
```

```
RMSE shifted 1 year:50400.987544437514
RMSE shifted 2 years:96285.5364206355
RMSE shifted 3 years:134135.72184545026
```

Dallas

In [47]:
```
baseline_graph('Dallas, TX')
```

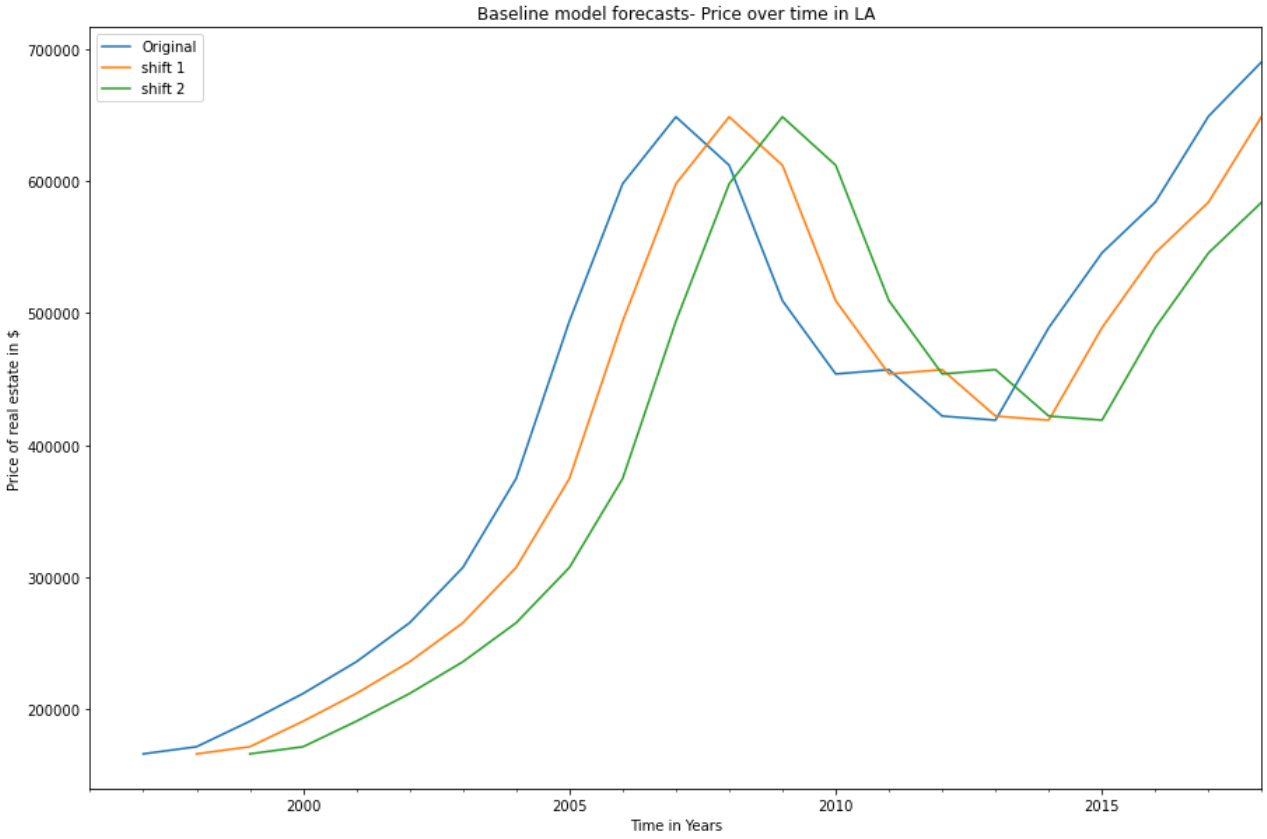


In [48]:
```
dallas_rmse = get_rmse('Dallas, TX')
baseline_rmse.append(dallas_rmse)
see_rmse('Dallas, TX')
```

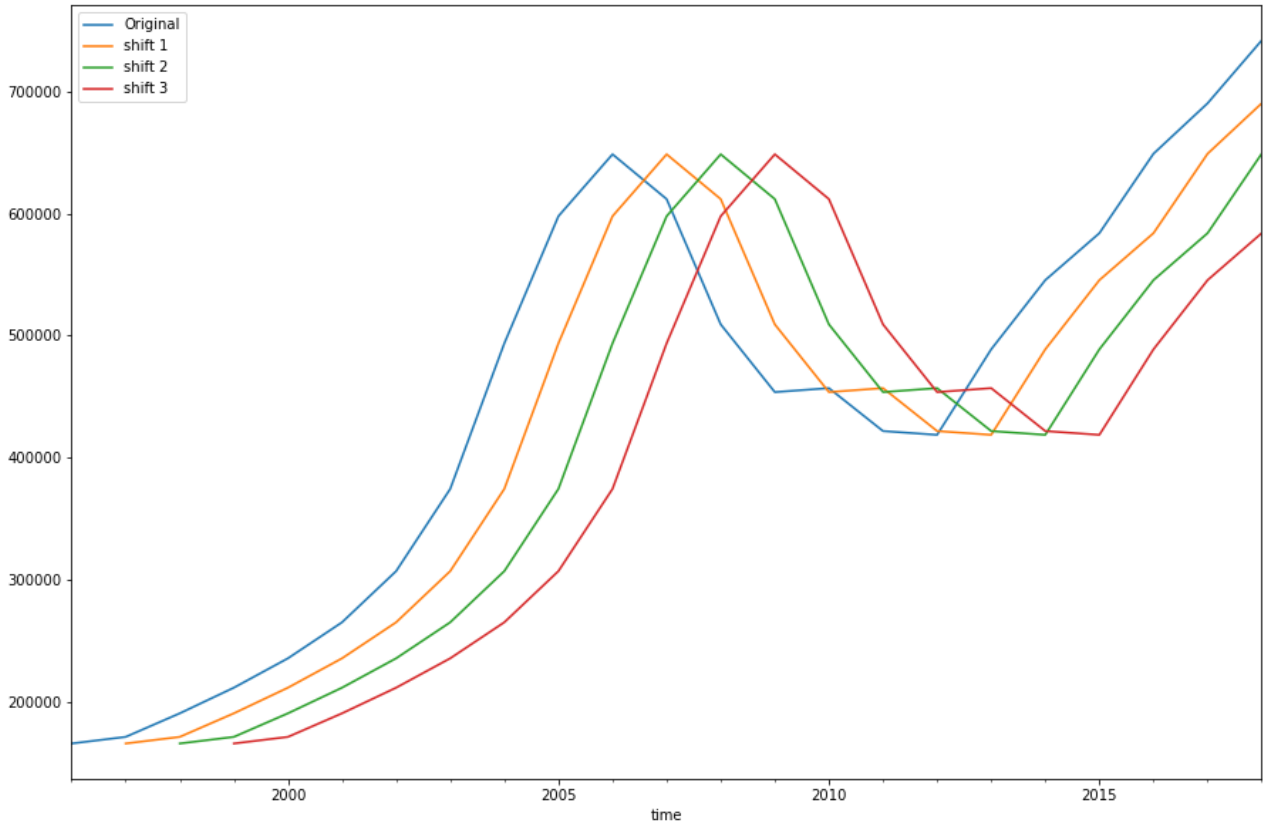
```
RMSE shifted 1 year:13330.761589101146
RMSE shifted 2 years:24214.157360244164
RMSE shifted 3 years:32990.72644319916
```

Los Angeles

In [49]:
```
# Saved the figure for illustration purpose in the ReadMe
fig, ax = plt.subplots(figsize=(15,10))
resampled_year['Los Angeles, CA'].shift(1).plot()
resampled_year['Los Angeles, CA'].shift(2).plot()
resampled_year['Los Angeles, CA'].shift(3).plot()
ax.legend(['Original', 'shift 1', 'shift 2', 'shift 3'])
ax.set_xlabel('Time in Years')
ax.set_ylabel('Price of real estate in $')
ax.set_title('Baseline model forecasts- Price over time in LA')
fig.savefig('figures/baselineLA.jpeg', dpi=500)
```



In [50]:
```
baseline_graph('Los Angeles, CA')
```

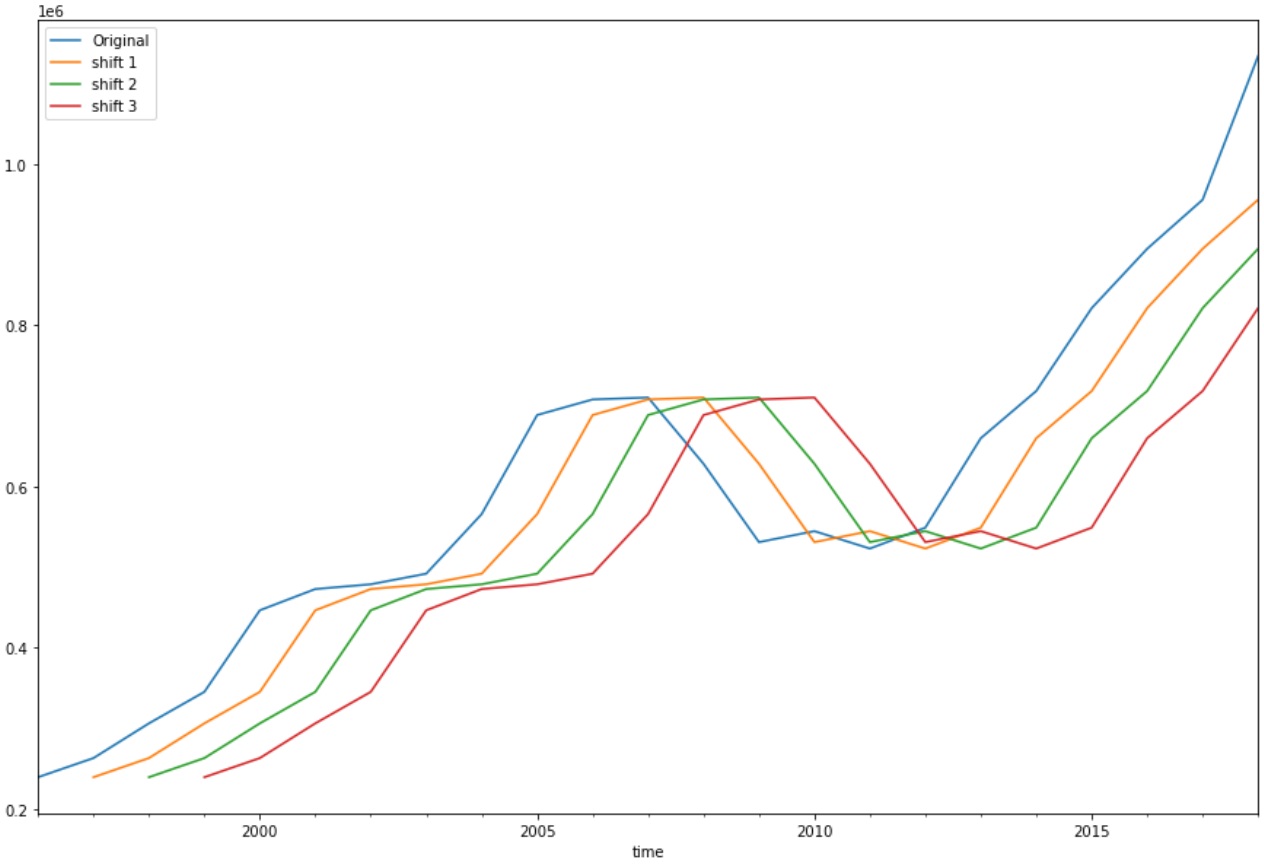


In [51]:
```
los_angeles_rmse = get_rmse('Los Angeles, CA')
baseline_rmse.append(los_angeles_rmse)
see_rmse('Los Angeles, CA')
```

```
RMSE shifted 1 year:56643.61975625242
RMSE shifted 2 years:107524.3184010903
RMSE shifted 3 years:151579.98012600475
```

San Jose

In [52]:
```
baseline_graph('San Jose, CA')
```

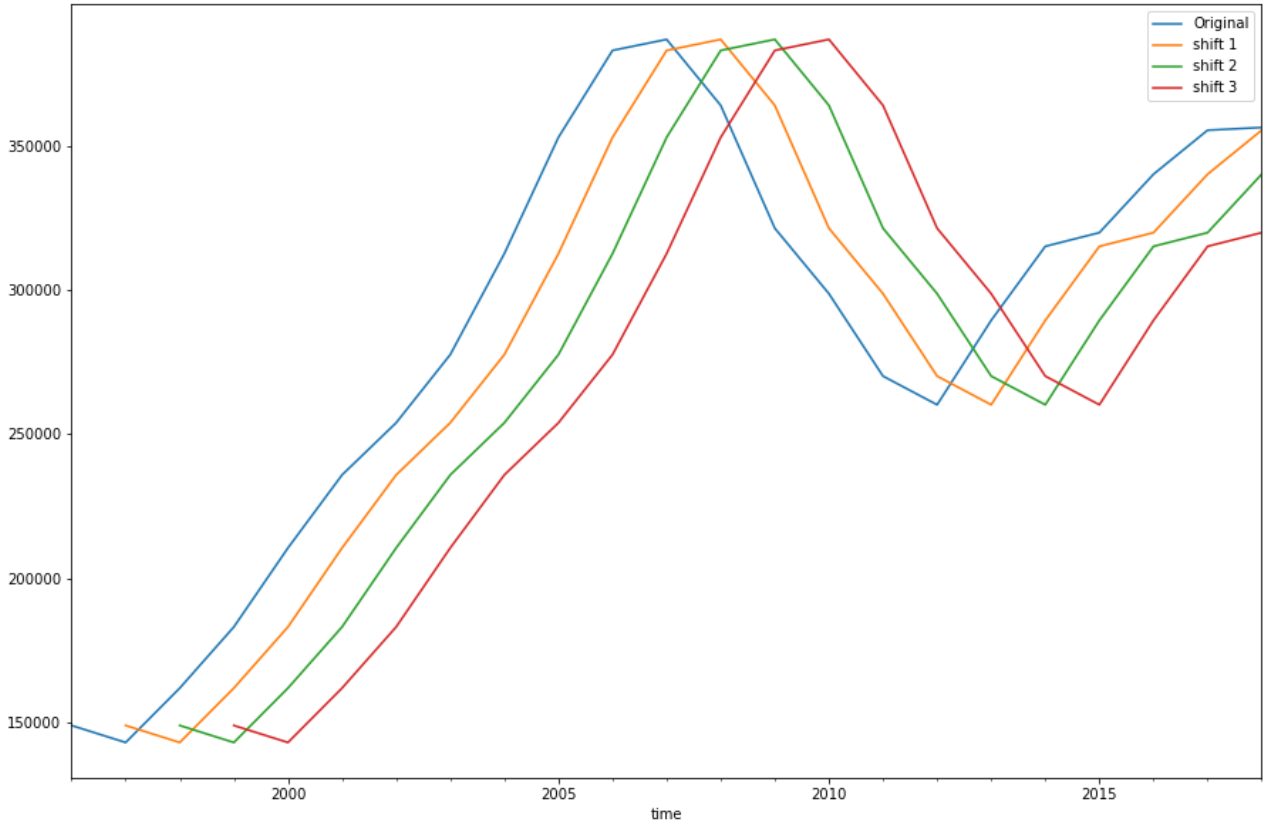


In [53]:
```
san_jose_rmse = get_rmse('San Jose, CA')
baseline_rmse.append(san_jose_rmse)
see_rmse('San Jose, CA')
```

```
RMSE shifted 1 year:74149.65748097192
RMSE shifted 2 years:127554.68528476419
RMSE shifted 3 years:176456.14079708306
```

Chicago

In [54]:
```
baseline_graph('Chicago, IL')
```

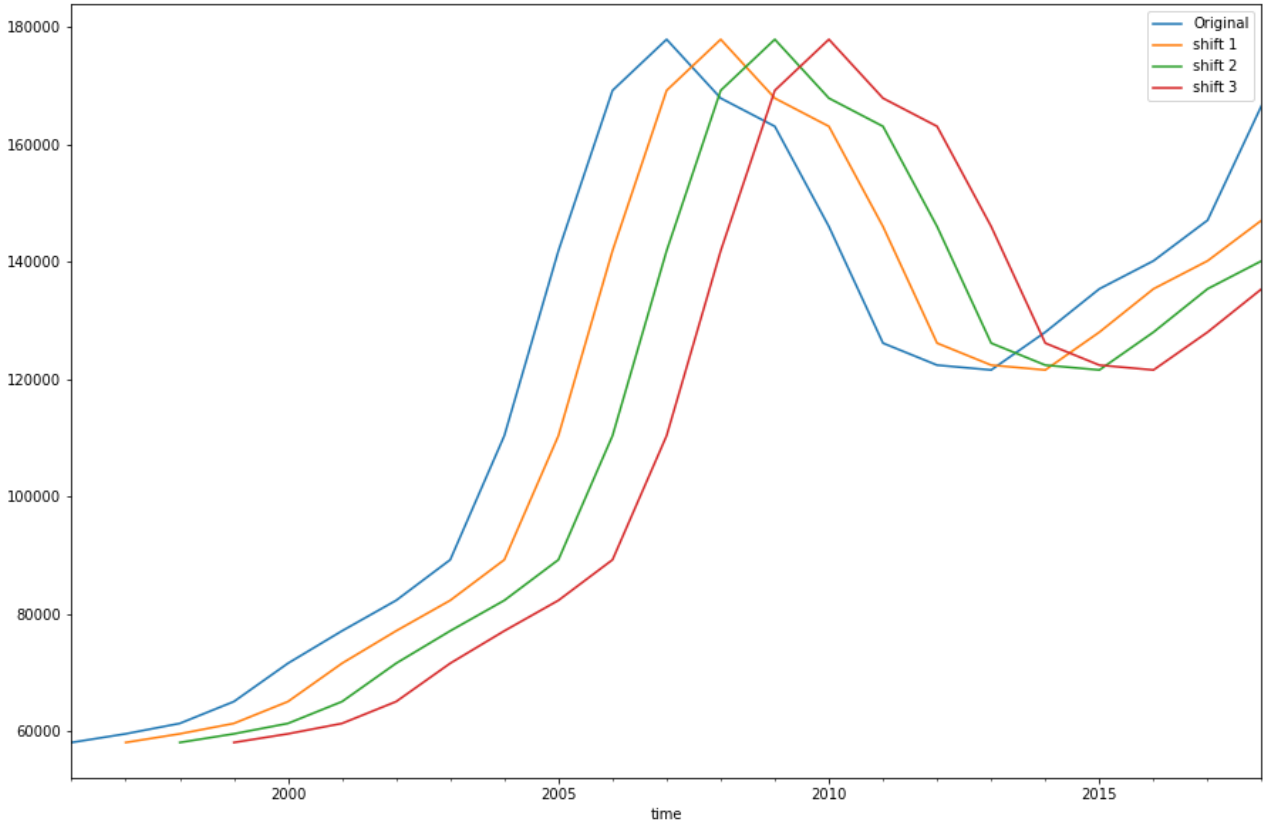


In [55]:
```
chicago_rmse = get_rmse('Chicago, IL')
baseline_rmse.append(chicago_rmse)
see_rmse('Chicago, IL')
```

```
RMSE shifted 1 year:24148.96831035532
RMSE shifted 2 years:46402.85872248657
RMSE shifted 3 years:66473.04575916467
```

Baltimore

In [56]:
```
baseline_graph('Baltimore, MD')
```

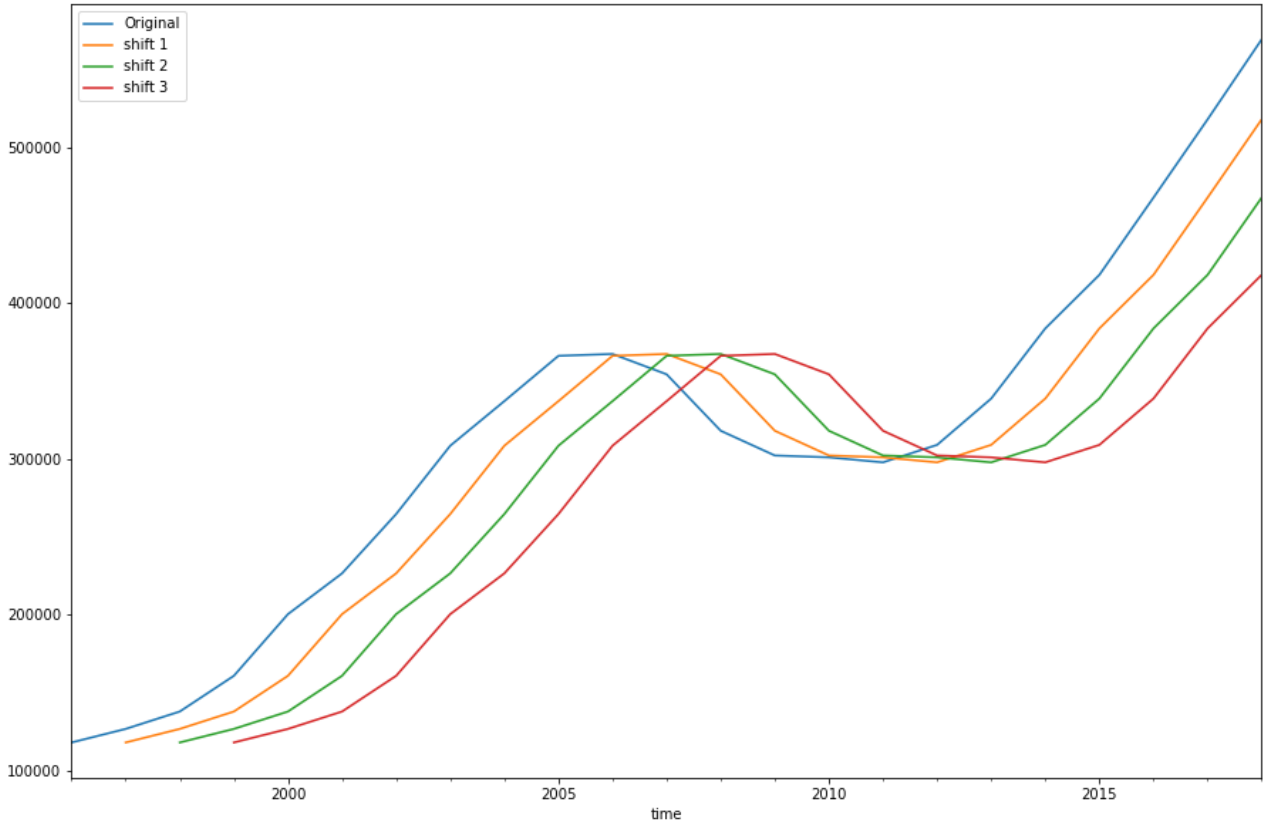


In [57]:
```
baltimore_rmse = get_rmse('Baltimore, MD')
baseline_rmse.append(baltimore_rmse)
see_rmse('Baltimore, MD')
```

```
RMSE shifted 1 year:13152.216370427665
RMSE shifted 2 years:24621.963192007486
RMSE shifted 3 years:34829.95298589994
```

Boston

In [58]:
```
baseline_graph('Boston, MA')
```



In [59]:
```
boston_rmse = get_rmse('Boston, MA')
baseline_rmse.append(boston_rmse)
see_rmse('Boston, MA')
```

```
RMSE shifted 1 year:31303.32223510063
RMSE shifted 2 years:60369.33893404278
RMSE shifted 3 years:87084.18753568297
```

In [60]:
```
# average rmse from the 10 cities
np.mean(baseline_rmse)
```

Out[60]:
```
118083.63296326711
```

ARIMA model

In [61]:
```
# used this helper function to give a frequency attribute to our data frame
def add_freq(idx, freq=None):
    """Add a frequency attribute to idx, through inference or directly.

    Returns a copy.  If `freq` is None, it is inferred.
    """

    idx = idx.copy()
    if freq is None:
        if idx.freq is None:
            freq = pd.infer_freq(idx)
        else:
            return idx
    idx.freq = pd.tseries.frequencies.to_offset(freq)
    if idx.freq is None:
        raise AttributeError('no discernible frequency found to `idx`.  Specify'
```

```
                                    ' a frequency string with `freq`.')
    return idx
```

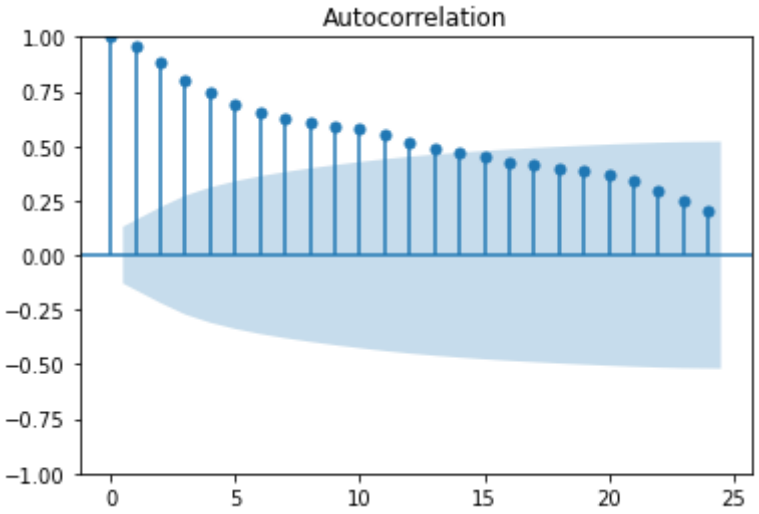
In [62]:
```
#We are adding frequency attribute to our dataframe index.
melted_df.index = add_freq(melted_df.index)
#Check the length of our dataframe and proportion of testing data
36 / len(melted_df)
```

Out[62]: 0.13584905660377358

We need to split the data into a training split and testing split. We intend our model to be able to predict house prices for the next three years, so it makes intuitive sense to make the testing data three years of prices. This will also allow give us a better idea of our performance on the data, as we won't be compared to markets further than three years in the future. We created a class that will fit an ARIMA model to a city's time series training data, summarize the model, make predictions for the testing data, plot those predictions versus the ground truth, and return a root mean squared error for the predictions. We ran through several different values in the ARIMA model and found that a p, d, q of 1, 2, 3 gave us the smallest combined root mean squared error for all the different cities. Our model ended up with an average RMSE of about $33,500 compared to the 118,000$ of the baseline model, a huge improvement.

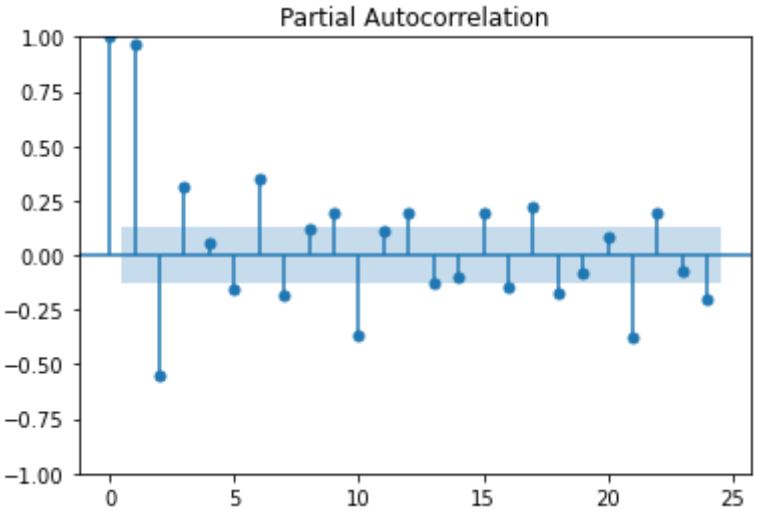
In [63]:
```
# make all but the last 3 years of data the training split
train = melted_df.iloc[:-36]
test = melted_df.iloc[-36:]
```

In [64]:
```
#Plotting the acf for Washington DC
plot_acf(train['Washington, DC'].diff().dropna());
```



In [65]:
```
#Plotted our PACF for Washington DC
plot_pacf(train['Washington, DC'].diff().dropna());
```

```
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/graphics/tsaplots.py:348: FutureWarning:

The default method 'yw' can produce PACF values outside of the [-1,1] interval. After 0.13, the default will change tounadjusted Yule-Walker ('ywm'). You can use this method now by setting method='ywm'.
```



In [66]:
```
class arima_mod:

    def __init__(self, city = 'Washington, DC'):
        self.city = city


    def model(self, df_train, df_test,p,d,q):
        #Fitting our model using ARIMA and instantiating it
        self.model_fit = ARIMA(df_train[self.city], order = [p,d,q]).fit()
        #Creating our prediction
        self.y_hat_test_ = self.model_fit.predict(start=df_test[self.city].index
                                                  end=df_test[self.city].index[-
        self.model_summary_ = self.model_fit.summary()
        self.rmse_ = mean_squared_error(df_test[self.city],
                                        self.y_hat_test_,
                                        squared=False)
        print(self.model_summary_)
        print('-'*23)
        print('-'*23)
        print(f'RMSE: {self.rmse_}')

    def plot(self, df_test):
        fig, ax = plt.subplots(figsize = (12,8))
        ax.plot(df_test[self.city])
        ax.plot(self.y_hat_test_)
        ax.legend(['Original', 'Predicted'])
        ax.set_title(f'Original vs Predicted home values for {self.city}')
```

In [67]:
```
# based on our acf and pacf graphs, it seems that 2, 1, 5 would be a good start
rmse_list = []
for city in city_list:
    city_model = arima_mod(city)
    city_model.model(train, test, 2,1,5)
    city_model.plot(test)
    rmse_list.append(city_model.rmse_)
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:          Washington, DC   No. Observations:                  229
Model:                 ARIMA(2, 1, 5)   Log Likelihood               -2816.379
Date:                Fri, 13 May 2022   AIC                           5648.758
Time:                        12:02:24   BIC                           5676.193
Sample:                    04-01-1996   HQIC                          5659.827
```

```
                           - 04-01-2015
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4499      5.249      0.086      0.932      -9.839      10.738
ar.L2          0.4827      5.014      0.096      0.923      -9.345      10.310
ma.L1         -0.2105      5.250     -0.040      0.968     -10.500      10.079
ma.L2         -0.3524      3.756     -0.094      0.925      -7.714       7.009
ma.L3         -0.0646      0.050     -1.284      0.199      -0.163       0.034
ma.L4         -0.1117      0.370     -0.302      0.763      -0.837       0.614
ma.L5         -0.0400      0.403     -0.099      0.921      -0.829       0.749
sigma2      5.348e+05      0.001   8.81e+08      0.000    5.35e+05    5.35e+05
==============================================================================
===
Ljung-Box (L1) (Q):                 134.57   Jarque-Bera (JB):              194
8.58
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               0.95   Skew:                           -
1.90
Prob(H) (two-sided):                  0.84   Kurtosis:                       1
6.81
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 4.1e+2
4. Standard errors may be unstable.
------------------------
------------------------
RMSE: 49394.25838149877
                               SARIMAX Results
==============================================================================
Dep. Variable:           New York, NY   No. Observations:                  229
Model:                ARIMA(2, 1, 5)   Log Likelihood               -2220.744
Date:               Fri, 13 May 2022   AIC                         4457.488
Time:                       12:02:25   BIC                         4484.923
Sample:                   04-01-1996   HQIC                        4468.557
                           - 04-01-2015
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.2059      0.594      2.030      0.042       0.041       2.370
ar.L2         -0.2443      0.569     -0.429      0.668      -1.360       0.871
ma.L1         -1.1471      0.595     -1.929      0.054      -2.312       0.018
ma.L2          0.2084      0.534      0.390      0.697      -0.839       1.256
ma.L3         -0.0091      0.019     -0.478      0.632      -0.046       0.028
ma.L4          0.0019      0.021      0.092      0.927      -0.038       0.042
ma.L5          0.0129      0.014      0.894      0.371      -0.015       0.041
sigma2      7.447e+06   7.93e-08    9.4e+13      0.000    7.45e+06    7.45e+06
==============================================================================
===
Ljung-Box (L1) (Q):                  92.77   Jarque-Bera (JB):                5
9.24
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               2.91   Skew:                           -
0.13
Prob(H) (two-sided):                  0.00   Kurtosis:
5.48
==============================================================================
```

```
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 1.04e+
31. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 93872.26139683934
                               SARIMAX Results
==========================================================================
Dep. Variable:      San Francisco, CA   No. Observations:                 229
Model:                 ARIMA(2, 1, 5)   Log Likelihood               -2721.162
Date:                Fri, 13 May 2022   AIC                           5458.324
Time:                        12:02:26   BIC                           5485.758
Sample:                    04-01-1996   HQIC                          5469.393
                         - 04-01-2015
Covariance Type:                  opg
==========================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------
ar.L1          0.0007      0.008      0.088      0.930      -0.015       0.016
ar.L2          0.9993      0.008    129.413      0.000       0.984       1.014
ma.L1          0.0892      0.019      4.607      0.000       0.051       0.127
ma.L2         -0.9338      0.011    -84.160      0.000      -0.956      -0.912
ma.L3         -0.0324      0.003    -10.091      0.000      -0.039      -0.026
ma.L4         -0.0652      0.003    -22.835      0.000      -0.071      -0.060
ma.L5         -0.0574      0.003    -17.898      0.000      -0.064      -0.051
sigma2       7.63e+06   2.39e-09    3.2e+15      0.000    7.63e+06    7.63e+06
==========================================================================
===
Ljung-Box (L1) (Q):                 160.19   Jarque-Bera (JB):                 2
0.96
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               2.82   Skew:                            -
0.39
Prob(H) (two-sided):                  0.00   Kurtosis:
4.26
==========================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 9.7e+2
9. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 120897.30461431197
                               SARIMAX Results
==========================================================================
Dep. Variable:            Seattle, WA   No. Observations:                 229
Model:                 ARIMA(2, 1, 5)   Log Likelihood               -2537.461
Date:                Fri, 13 May 2022   AIC                           5090.922
Time:                        12:02:27   BIC                           5118.356
Sample:                    04-01-1996   HQIC                          5101.991
                         - 04-01-2015
Covariance Type:                  opg
==========================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------
ar.L1          1.9134      0.020     97.534      0.000       1.875       1.952
ar.L2         -0.9160      0.018    -49.832      0.000      -0.952      -0.880
```

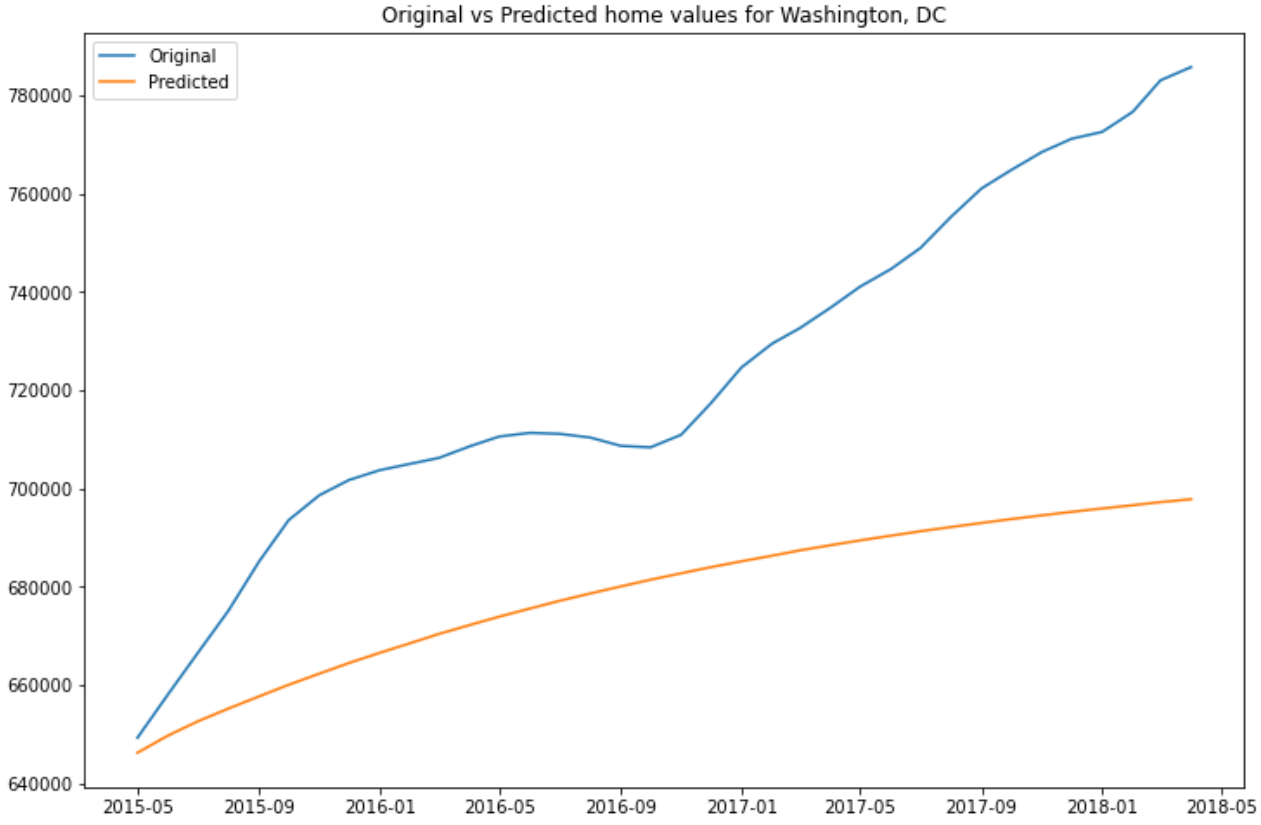
```
ma.L1         -1.8507      0.020    -94.548      0.000      -1.889      -1.812
ma.L2          0.8718      0.017     50.423      0.000       0.838       0.906
ma.L3         -0.0073      0.003     -2.446      0.014      -0.013      -0.001
ma.L4         -0.0150      0.003     -4.648      0.000      -0.021      -0.009
ma.L5          0.0074      0.002      3.466      0.001       0.003       0.012
sigma2      8.713e+05   5.48e-09   1.59e+14      0.000    8.71e+05    8.71e+05
===================================================================================
===
Ljung-Box (L1) (Q):                 176.41   Jarque-Bera (JB):                 8
0.87
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.57   Skew:                            -
1.10
Prob(H) (two-sided):                  0.05   Kurtosis:
4.91
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 2.83e+
30. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 141561.18748246835
                                SARIMAX Results
===================================================================================
Dep. Variable:             Dallas, TX   No. Observations:                   229
Model:                 ARIMA(2, 1, 5)   Log Likelihood               -1945.752
Date:                Fri, 13 May 2022   AIC                           3907.503
Time:                        12:02:27   BIC                           3934.938
Sample:                    04-01-1996   HQIC                          3918.572
                         - 04-01-2015
Covariance Type:                  opg
===================================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
ar.L1           0.9218     24.478      0.038      0.970     -47.054      48.897
ar.L2          -0.2645     12.312     -0.021      0.983     -24.396      23.867
ma.L1          -0.8979     24.478     -0.037      0.971     -48.874      47.078
ma.L2           0.2605     11.730      0.022      0.982     -22.729      23.250
ma.L3           0.0015      0.157      0.010      0.992      -0.307       0.310
ma.L4           0.0031      0.077      0.040      0.968      -0.148       0.154
ma.L5           0.0002      0.107      0.002      0.998      -0.209       0.209
sigma2      1.277e+06   2.31e+04     55.311      0.000    1.23e+06    1.32e+06
===================================================================================
===
Ljung-Box (L1) (Q):                  28.76   Jarque-Bera (JB):             4380
7.69
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              11.01   Skew:                            -
5.80
Prob(H) (two-sided):                  0.00   Kurtosis:                         6
9.91
===================================================================================
===

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

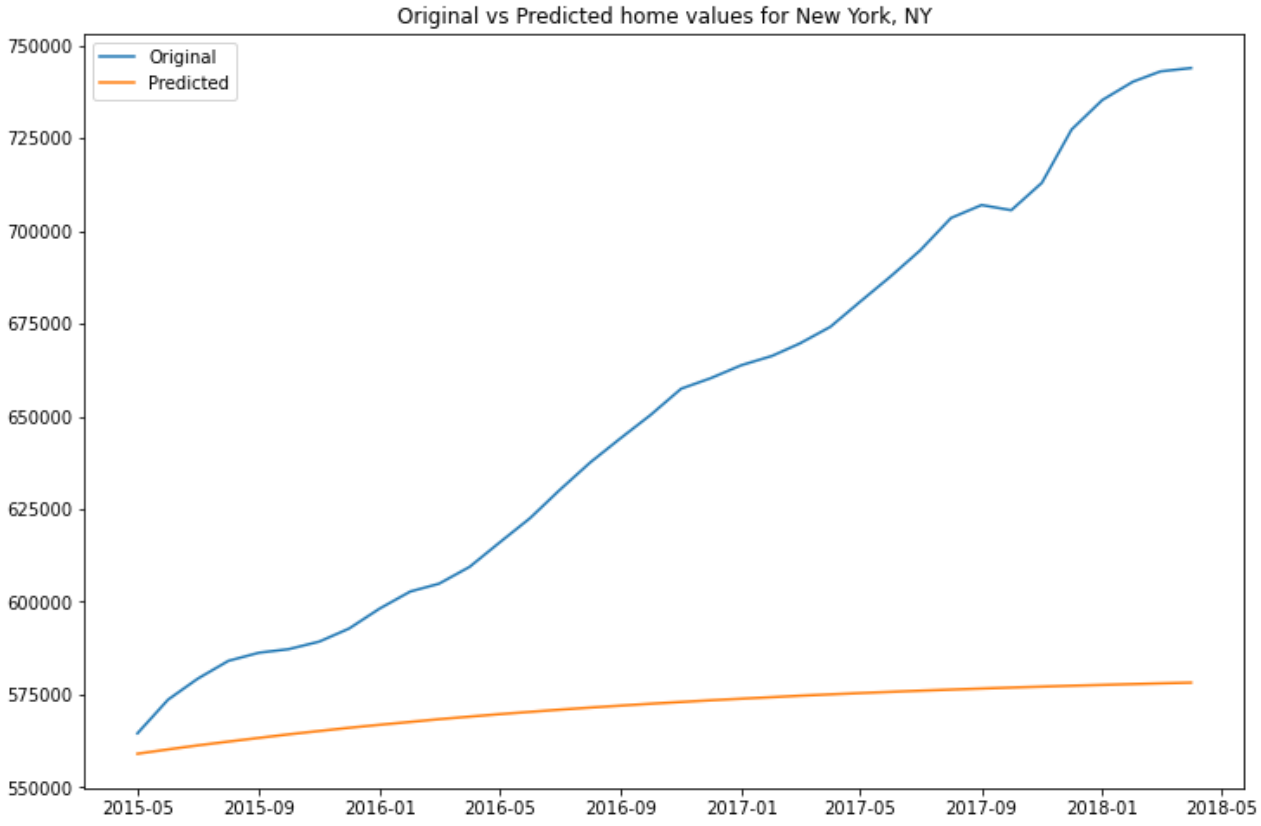
```
RMSE: 54145.25507618063
                               SARIMAX Results
==============================================================================
Dep. Variable:        Los Angeles, CA   No. Observations:                  229
Model:                 ARIMA(2, 1, 5)   Log Likelihood               -2306.723
Date:                Fri, 13 May 2022   AIC                           4629.446
Time:                        12:02:28   BIC                           4656.880
Sample:                    04-01-1996   HQIC                          4640.515
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.9787      0.004    518.328      0.000       1.971       1.986
ar.L2         -0.9826      0.004   -255.981      0.000      -0.990      -0.975
ma.L1         -1.8701      0.005   -410.416      0.000      -1.879      -1.861
ma.L2          0.8438      0.007    126.539      0.000       0.831       0.857
ma.L3          0.0390      0.008      5.126      0.000       0.024       0.054
ma.L4         -0.0732      0.010     -7.436      0.000      -0.093      -0.054
ma.L5          0.0669      0.007      9.280      0.000       0.053       0.081
sigma2      5.106e+06   2.65e-10   1.93e+16      0.000    5.11e+06    5.11e+06
==============================================================================
===
Ljung-Box (L1) (Q):                 122.17   Jarque-Bera (JB):               14
8.24
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.00   Skew:                            -
0.85
Prob(H) (two-sided):                  1.00   Kurtosis:
6.57
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 3.02e+
31. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 101012.39183890459
                               SARIMAX Results
==============================================================================
Dep. Variable:          San Jose, CA   No. Observations:                  229
Model:                 ARIMA(2, 1, 5)   Log Likelihood               -2526.787
Date:                Fri, 13 May 2022   AIC                           5069.573
Time:                        12:02:30   BIC                           5097.008
Sample:                    04-01-1996   HQIC                          5080.642
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.9752      0.006    305.133      0.000       1.962       1.988
ar.L2         -0.9755      0.006   -155.359      0.000      -0.988      -0.963
ma.L1         -1.8853      0.007   -283.250      0.000      -1.898      -1.872
ma.L2          0.8263      0.008    107.089      0.000       0.811       0.841
ma.L3          0.0923      0.008     12.037      0.000       0.077       0.107
ma.L4         -0.0731      0.006    -12.849      0.000      -0.084      -0.062
ma.L5          0.0406      0.004     11.529      0.000       0.034       0.047
sigma2      4.423e+06   3.21e-10   1.38e+16      0.000    4.42e+06    4.42e+06
==============================================================================
===
Ljung-Box (L1) (Q):                 166.09   Jarque-Bera (JB):                1
```

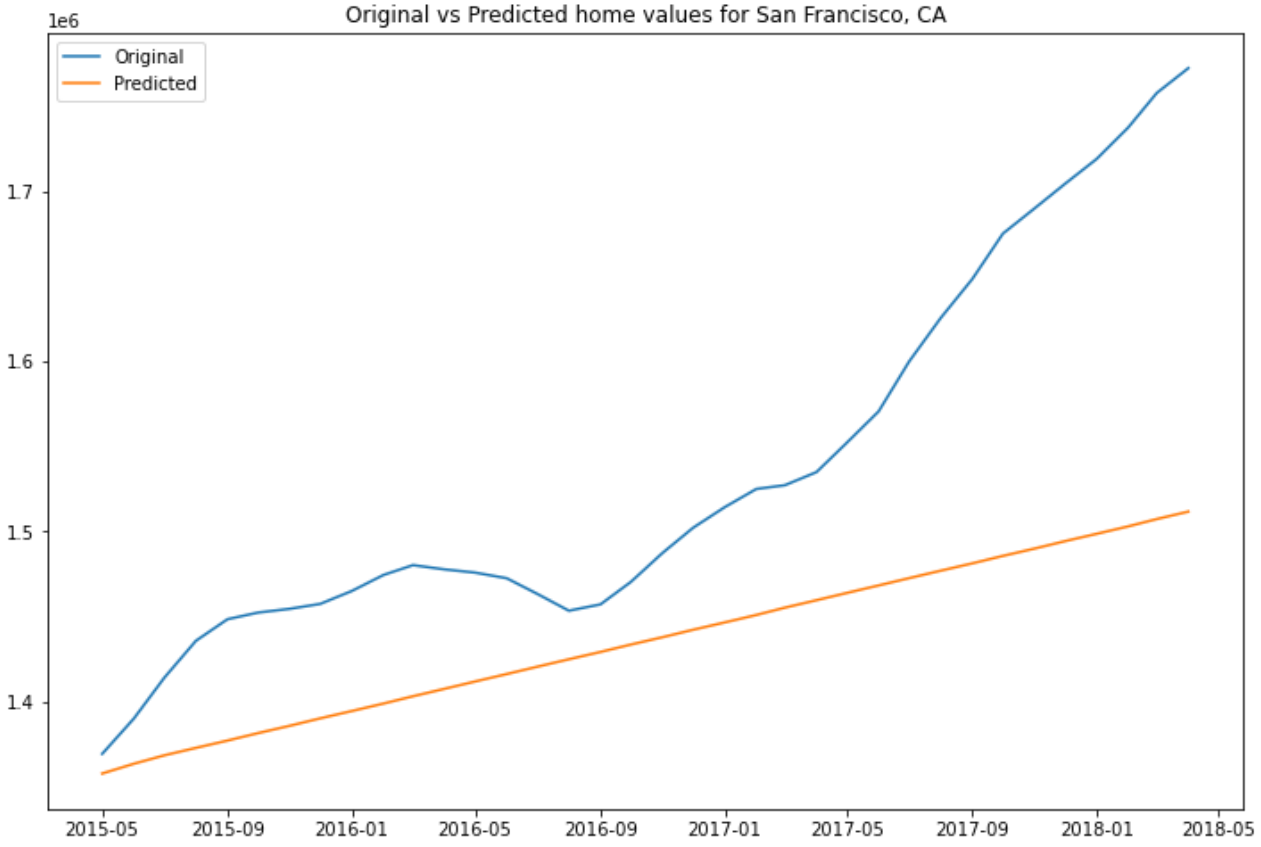
```
6.56
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               2.21   Skew:                             -
0.15
Prob(H) (two-sided):                  0.00   Kurtosis:
4.29
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 2.11e+
31. Standard errors may be unstable.
------------------------
------------------------
RMSE: 126428.52165092973
                               SARIMAX Results
================================================================================
Dep. Variable:            Chicago, IL   No. Observations:                   229
Model:                 ARIMA(2, 1, 5)   Log Likelihood               -2072.821
Date:                Fri, 13 May 2022   AIC                            4161.642
Time:                        12:02:31   BIC                            4189.077
Sample:                    04-01-1996   HQIC                           4172.711
                         - 04-01-2015
Covariance Type:                  opg
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
ar.L1          1.9500      0.047     41.346      0.000       1.858       2.042
ar.L2         -0.9514      0.045    -21.011      0.000      -1.040      -0.863
ma.L1         -1.9051      0.047    -40.445      0.000      -1.997      -1.813
ma.L2          0.8967      0.044     20.161      0.000       0.810       0.984
ma.L3          0.0074      0.015      0.488      0.625      -0.022       0.037
ma.L4         -0.0059      0.017     -0.356      0.722      -0.039       0.027
ma.L5          0.0092      0.009      0.970      0.332      -0.009       0.028
sigma2      3.893e+06   1.36e-08   2.86e+14      0.000    3.89e+06    3.89e+06
================================================================================
===
Ljung-Box (L1) (Q):                 180.85   Jarque-Bera (JB):                 6
6.99
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.72   Skew:                             -
0.05
Prob(H) (two-sided):                  0.02   Kurtosis:
5.65
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 9.63e+
29. Standard errors may be unstable.
------------------------
------------------------
RMSE: 18222.667259101647
                               SARIMAX Results
================================================================================
Dep. Variable:           Baltimore, MD   No. Observations:                   229
Model:                 ARIMA(2, 1, 5)   Log Likelihood               -1882.176
Date:                Fri, 13 May 2022   AIC                            3780.352
Time:                        12:02:31   BIC                            3807.787
```

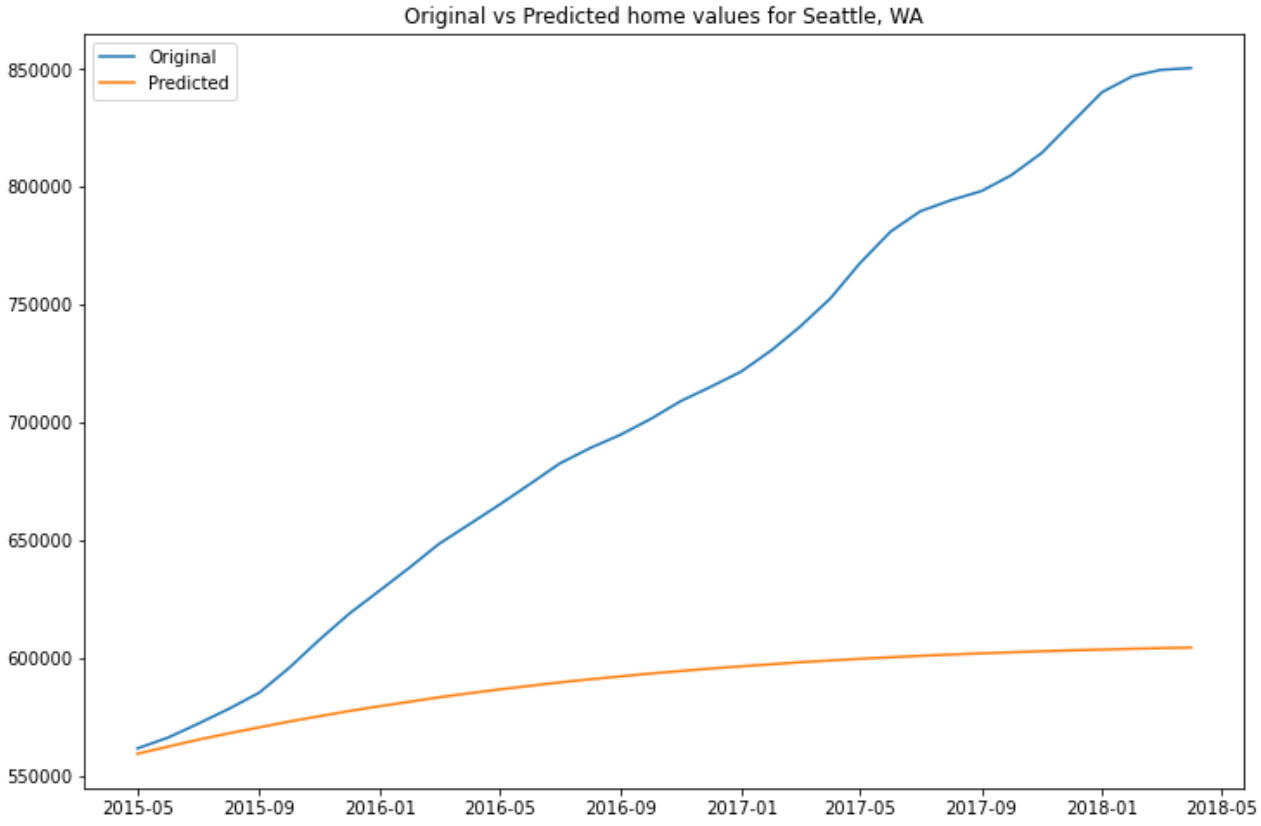
```
Sample:                     04-01-1996   HQIC                          3791.421
                          - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.5206      3.942      0.386      0.700      -6.206       9.247
ar.L2         -0.5489      3.682     -0.149      0.881      -7.765       6.667
ma.L1         -1.3419      3.943     -0.340      0.734      -9.069       6.386
ma.L2          0.3715      2.977      0.125      0.901      -5.464       6.207
ma.L3          0.0377      0.287      0.132      0.895      -0.524       0.599
ma.L4         -0.0012      0.039     -0.032      0.975      -0.078       0.075
ma.L5          0.0049      0.023      0.211      0.833      -0.041       0.051
sigma2      5.998e+05      0.000   4.72e+09      0.000       6e+05       6e+05
===================================================================================
===
Ljung-Box (L1) (Q):                 115.99   Jarque-Bera (JB):               25
5.70
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               2.02   Skew:                            -
1.02
Prob(H) (two-sided):                  0.00   Kurtosis:
7.77
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 2.18e+
25. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 9705.316833039313
                               SARIMAX Results
==============================================================================
Dep. Variable:             Boston, MA   No. Observations:                  229
Model:                 ARIMA(2, 1, 5)   Log Likelihood               -2093.561
Date:                Fri, 13 May 2022   AIC                           4203.122
Time:                        12:02:32   BIC                           4230.557
Sample:                     04-01-1996   HQIC                          4214.191
                          - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.9515      0.030     64.193      0.000       1.892       2.011
ar.L2         -0.9530      0.029    -32.555      0.000      -1.010      -0.896
ma.L1         -1.8832      0.031    -60.267      0.000      -1.944      -1.822
ma.L2          0.8644      0.030     28.671      0.000       0.805       0.924
ma.L3          0.0197      0.017      1.182      0.237      -0.013       0.052
ma.L4         -0.0128      0.019     -0.681      0.496      -0.050       0.024
ma.L5          0.0149      0.011      1.332      0.183      -0.007       0.037
sigma2      2.827e+06   6.19e-09   4.57e+14      0.000    2.83e+06    2.83e+06
===================================================================================
===
Ljung-Box (L1) (Q):                 103.28   Jarque-Bera (JB):               16
0.09
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.85   Skew:
0.38
Prob(H) (two-sided):                  0.01   Kurtosis:
7.03
```

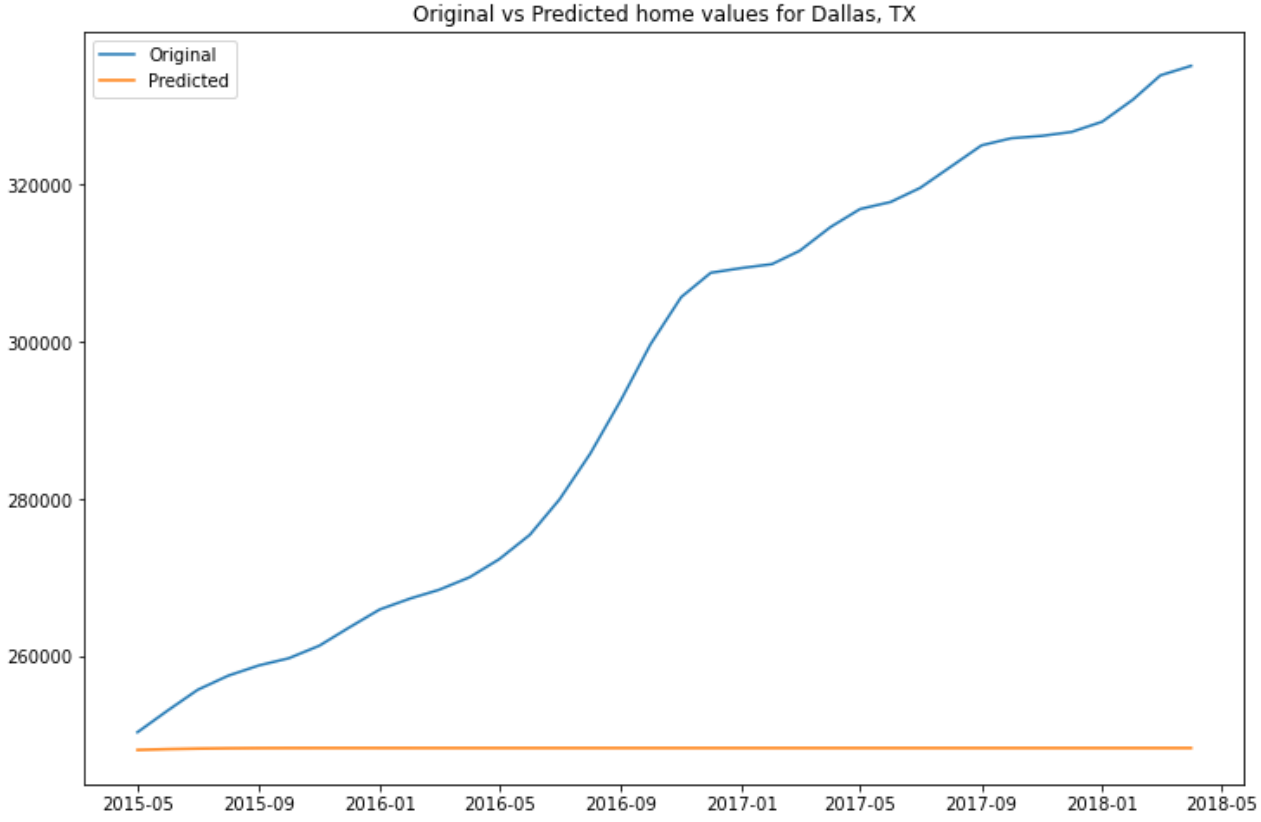
```
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 6.89e+
29. Standard errors may be unstable.
------------------------
------------------------
RMSE: 70922.40215618946
```

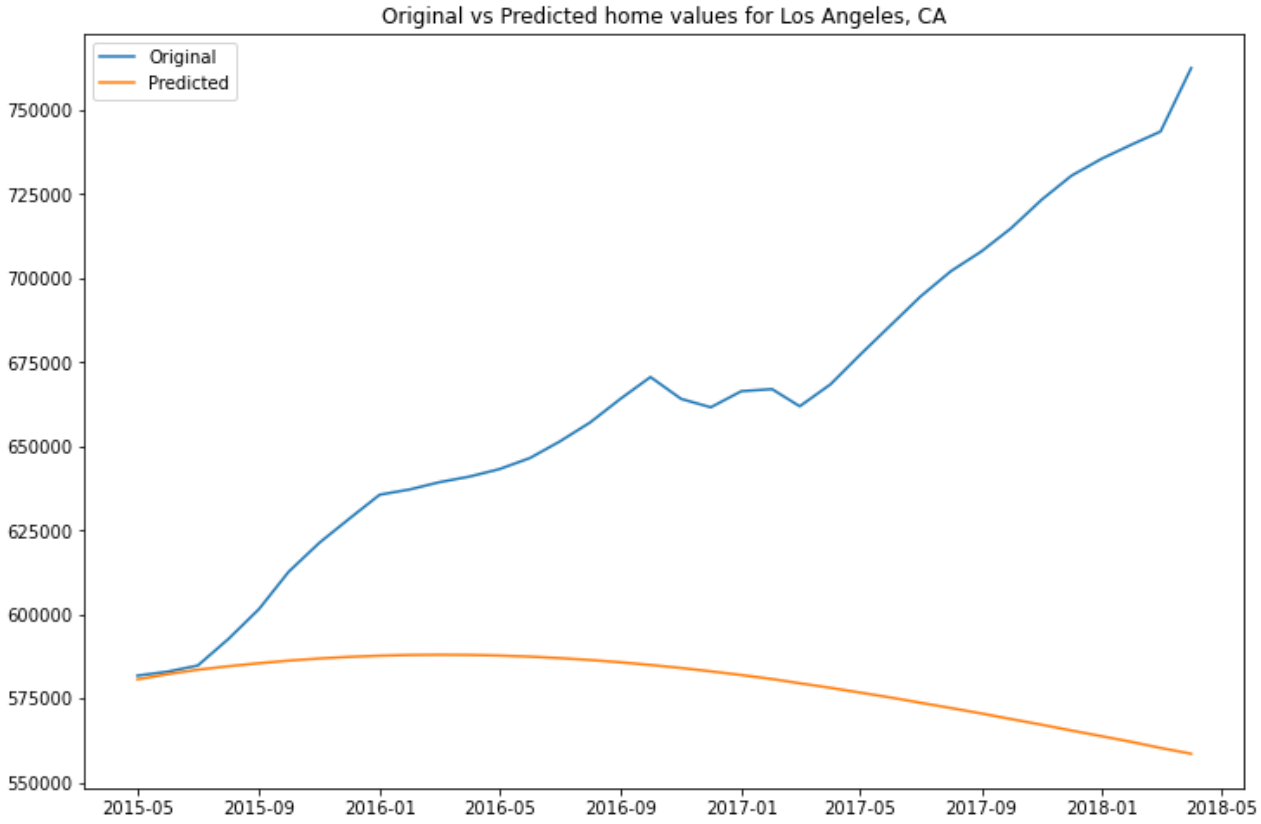


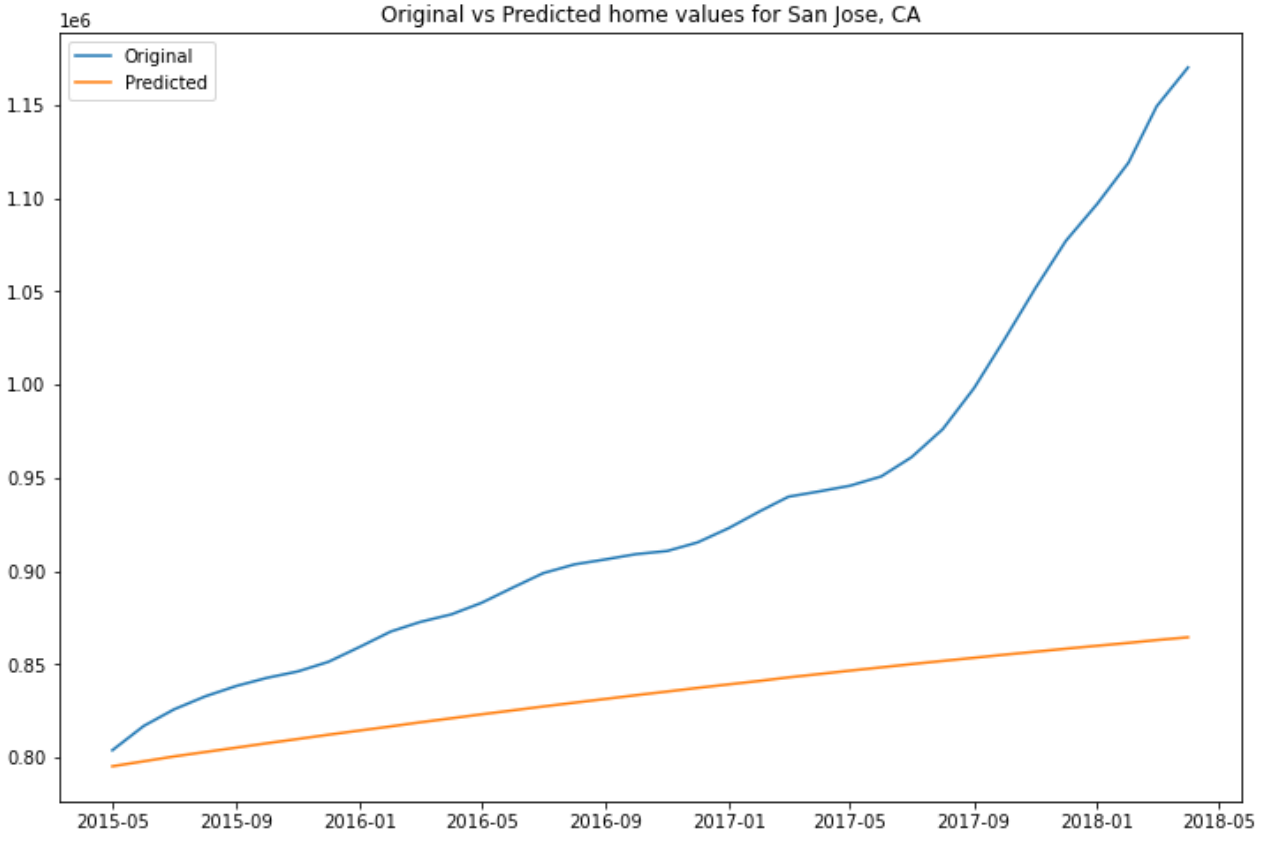

Original vs Predicted home values for New York, NY
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
550000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

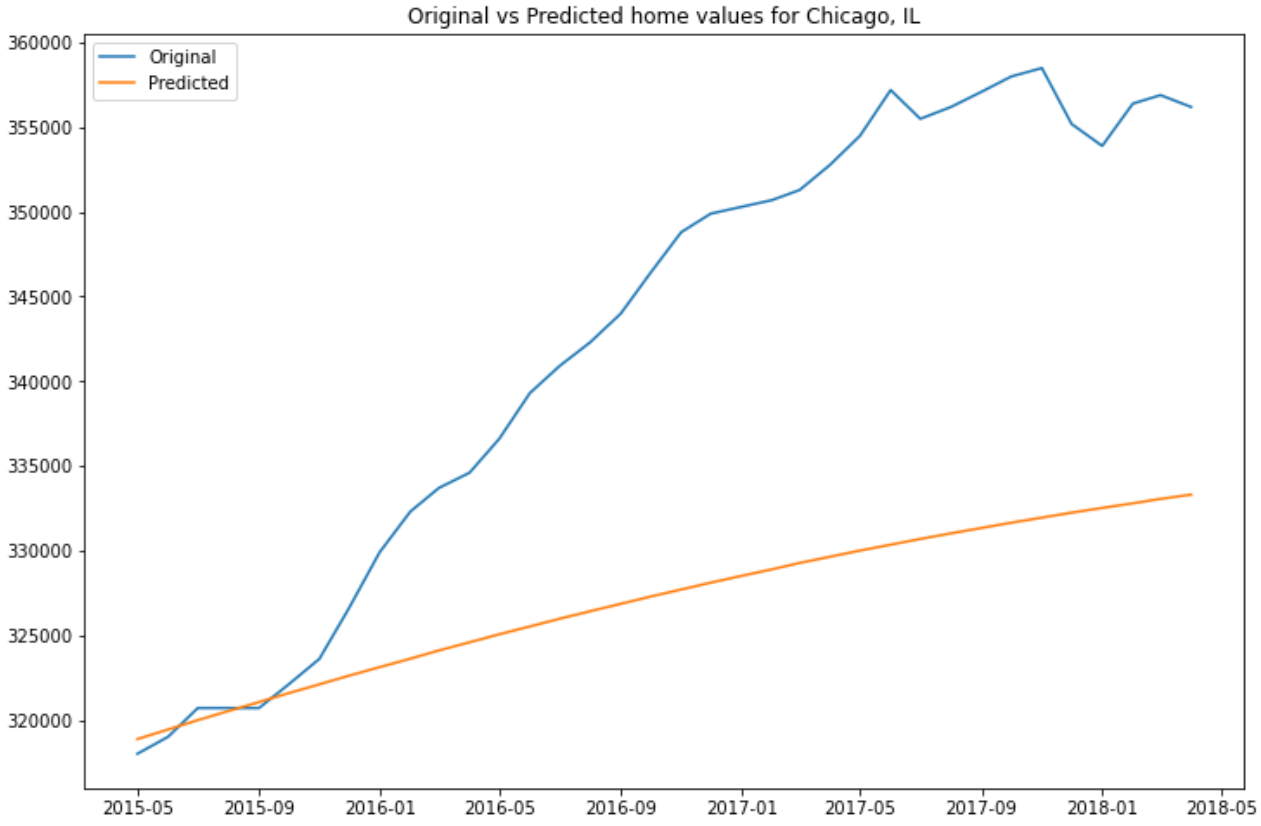

Original vs Predicted home values for San Francisco, CA
1e6
Original
Predicted
1.7
1.6
1.5
1.4
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

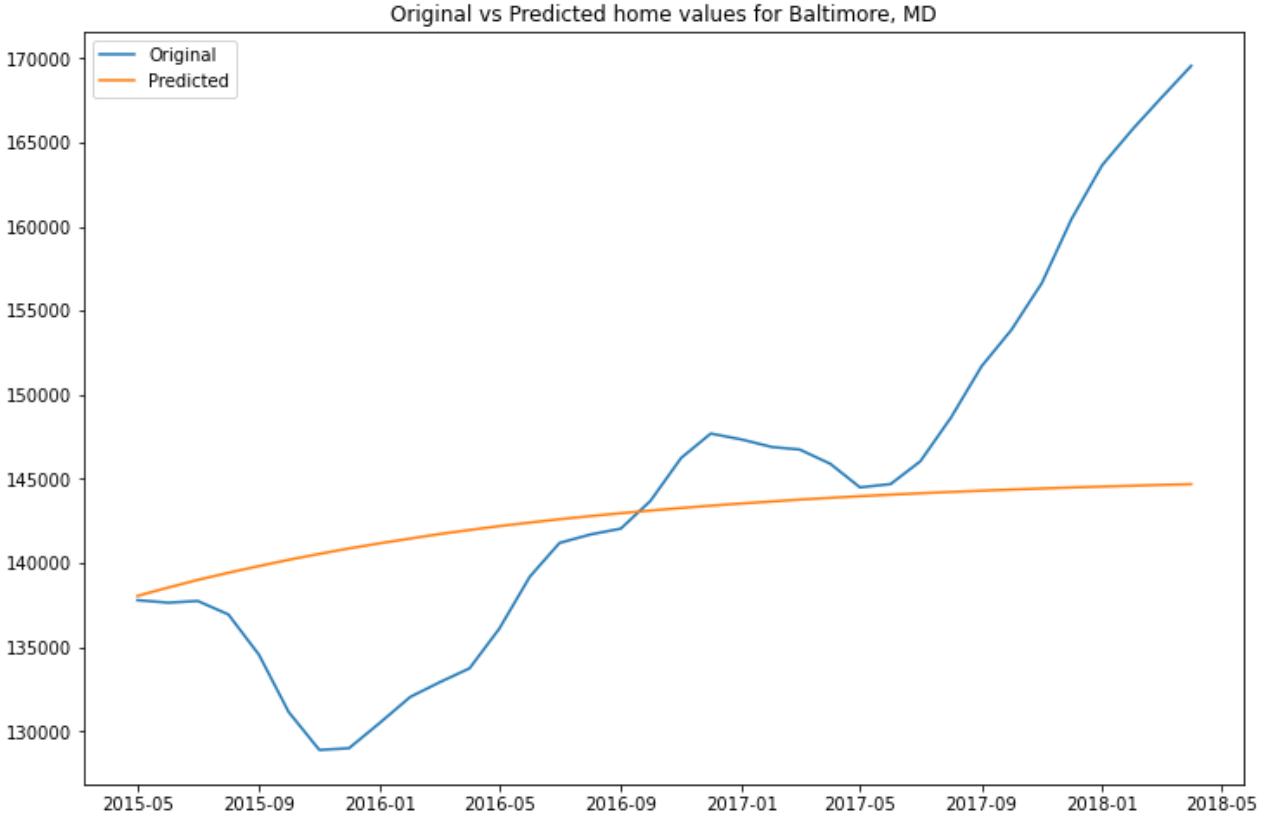

Original vs Predicted home values for Seattle, WA
Original
Predicted
850000
800000
750000
700000
650000
600000
550000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

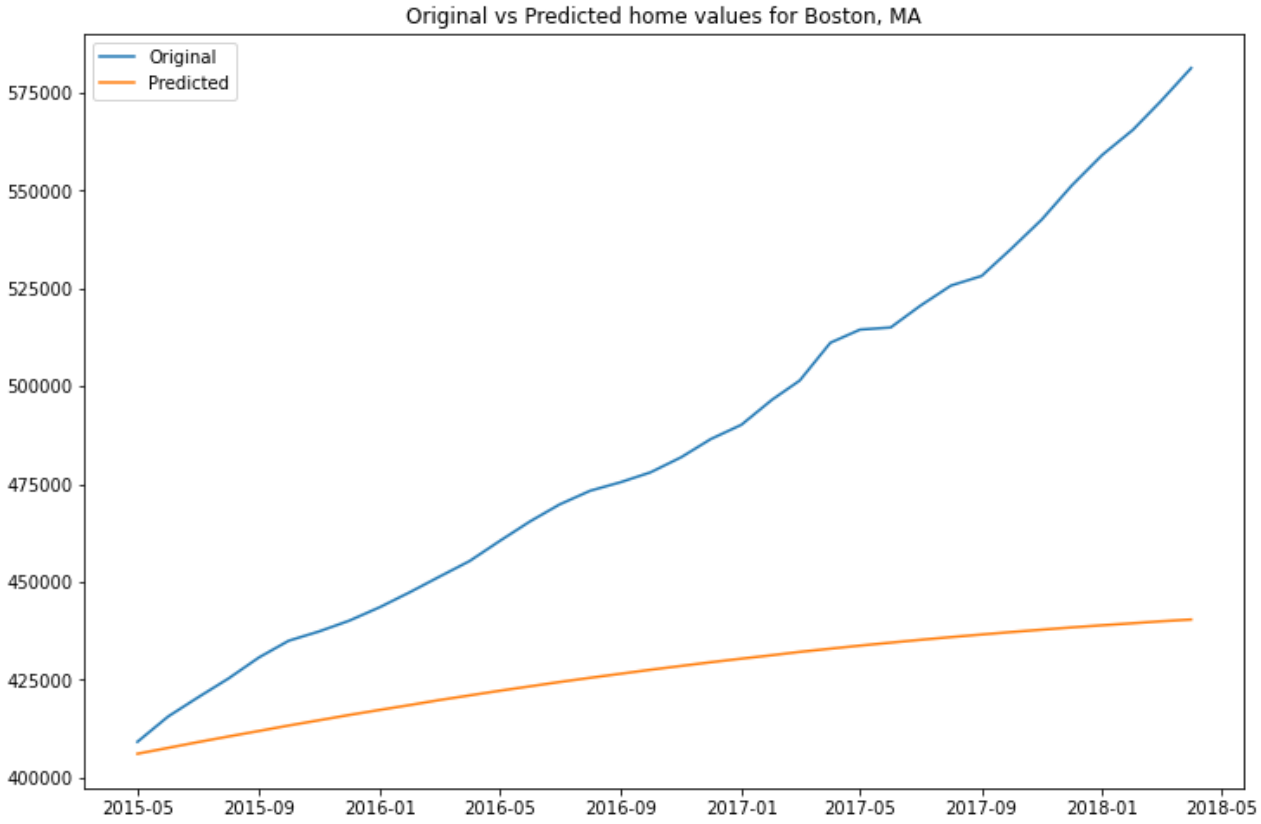

Original vs Predicted home values for Dallas, TX
Original
Predicted
320000
300000
280000
260000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Los Angeles, CA
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
550000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for San Jose, CA
1e6
Original
Predicted
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Chicago, IL
Original
Predicted
360000
355000
350000
345000
340000
335000
330000
325000
320000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Baltimore, MD
Original
Predicted
170000
165000
160000
155000
150000
145000
140000
135000
130000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05



In [68]:
```
# the mean of the rmse for every city
np.mean(rmse_list)
```

Out[68]: 78616.15666894638

In [69]:
```
# The previous graphs were pretty rough, and the rmse was as well, something mor
rmse_list = []
for city in city_list:
    city_model = arima_mod(city)
    city_model.model(train, test, 1, 1, 1)
    city_model.plot(test)
    rmse_list.append(city_model.rmse_)
```

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                 Washington, DC   No. Observations:                  229
Model:                 ARIMA(1, 1, 1)   Log Likelihood               -2141.888
Date:                Fri, 13 May 2022   AIC                           4289.777
Time:                        12:02:36   BIC                           4300.065
Sample:                    04-01-1996   HQIC                          4293.928
                         - 04-01-2015
Covariance Type:                  opg
==========================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9498      0.015     62.956      0.000       0.920       0.979
ma.L1         -0.8895      0.021    -41.807      0.000      -0.931      -0.848
sigma2      8.028e+06   4.19e-10   1.92e+16      0.000    8.03e+06    8.03e+06
==========================================================================================
===
Ljung-Box (L1) (Q):                 191.27   Jarque-Bera (JB):                 2
2.15
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.26   Skew:                             -
```

```
0.35
Prob(H) (two-sided):                  0.31   Kurtosis:
4.35
===============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 7.25e+
31. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 55752.977663068355
                               SARIMAX Results
===============================================================================
Dep. Variable:           New York, NY   No. Observations:                  229
Model:                 ARIMA(1, 1, 1)   Log Likelihood               -2197.813
Date:                Fri, 13 May 2022   AIC                         4401.626
Time:                        12:02:36   BIC                         4411.914
Sample:                  04-01-1996   HQIC                        4405.777
                       - 04-01-2015
Covariance Type:                 opg
===============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
ar.L1          0.9527      0.022     43.387      0.000       0.910       0.996
ma.L1         -0.9186      0.028    -33.282      0.000      -0.973      -0.864
sigma2      1.362e+07   2.79e-10   4.88e+16      0.000    1.36e+07    1.36e+07
===============================================================================
===
Ljung-Box (L1) (Q):                 100.00   Jarque-Bera (JB):                5
0.72
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               3.18   Skew:                            -
0.07
Prob(H) (two-sided):                  0.00   Kurtosis:
5.31
===============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 4.66e+
32. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 94205.75364498292
                               SARIMAX Results
===============================================================================
Dep. Variable:      San Francisco, CA   No. Observations:                  229
Model:                 ARIMA(1, 1, 1)   Log Likelihood               -2356.683
Date:                Fri, 13 May 2022   AIC                         4719.365
Time:                        12:02:37   BIC                         4729.653
Sample:                  04-01-1996   HQIC                        4723.516
                       - 04-01-2015
Covariance Type:                 opg
===============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
ar.L1          0.9999      0.001    761.367      0.000       0.997       1.002
ma.L1         -0.9981      0.011    -88.177      0.000      -1.020      -0.976
sigma2      5.539e+07   3.12e-12   1.78e+19      0.000    5.54e+07    5.54e+07
```

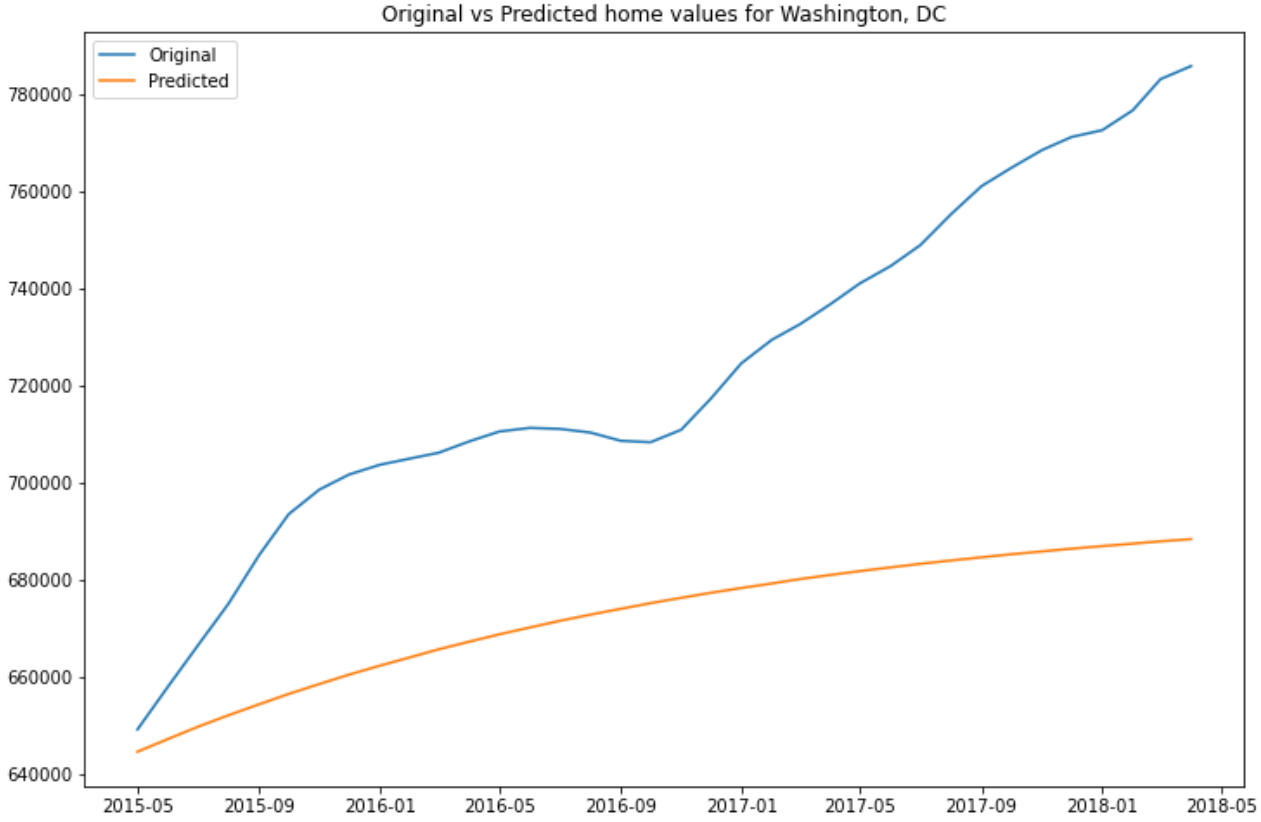
```
==========================================================================
===
Ljung-Box (L1) (Q):                 190.71   Jarque-Bera (JB):
0.06
Prob(Q):                              0.00   Prob(JB):
0.97
Heteroskedasticity (H):               8.89   Skew:
0.03
Prob(H) (two-sided):                  0.00   Kurtosis:
2.95
==========================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 1.97e+
35. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 120265.5068634642
                           SARIMAX Results
==========================================================================
Dep. Variable:            Seattle, WA   No. Observations:                 229
Model:                 ARIMA(1, 1, 1)   Log Likelihood               -2126.276
Date:                Fri, 13 May 2022   AIC                           4258.551
Time:                        12:02:37   BIC                           4268.839
Sample:                    04-01-1996   HQIC                          4262.702
                         - 04-01-2015
Covariance Type:                 opg
==========================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------
ar.L1          0.9531      0.015     63.399      0.000       0.924       0.983
ma.L1         -0.9134      0.019    -48.283      0.000      -0.951      -0.876
sigma2      7.071e+06   1.17e-10   6.03e+16      0.000    7.07e+06    7.07e+06
==========================================================================
===
Ljung-Box (L1) (Q):                 189.80   Jarque-Bera (JB):                2
0.63
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               2.49   Skew:                            -
0.71
Prob(H) (two-sided):                  0.00   Kurtosis:
3.39
==========================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 9.9e+3
33. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 150768.12793117162
                           SARIMAX Results
==========================================================================
Dep. Variable:              Dallas, TX   No. Observations:                 229
Model:                 ARIMA(1, 1, 1)   Log Likelihood               -1946.922
Date:                Fri, 13 May 2022   AIC                           3899.844
Time:                        12:02:38   BIC                           3910.132
Sample:                    04-01-1996   HQIC                          3903.994
                         - 04-01-2015
```

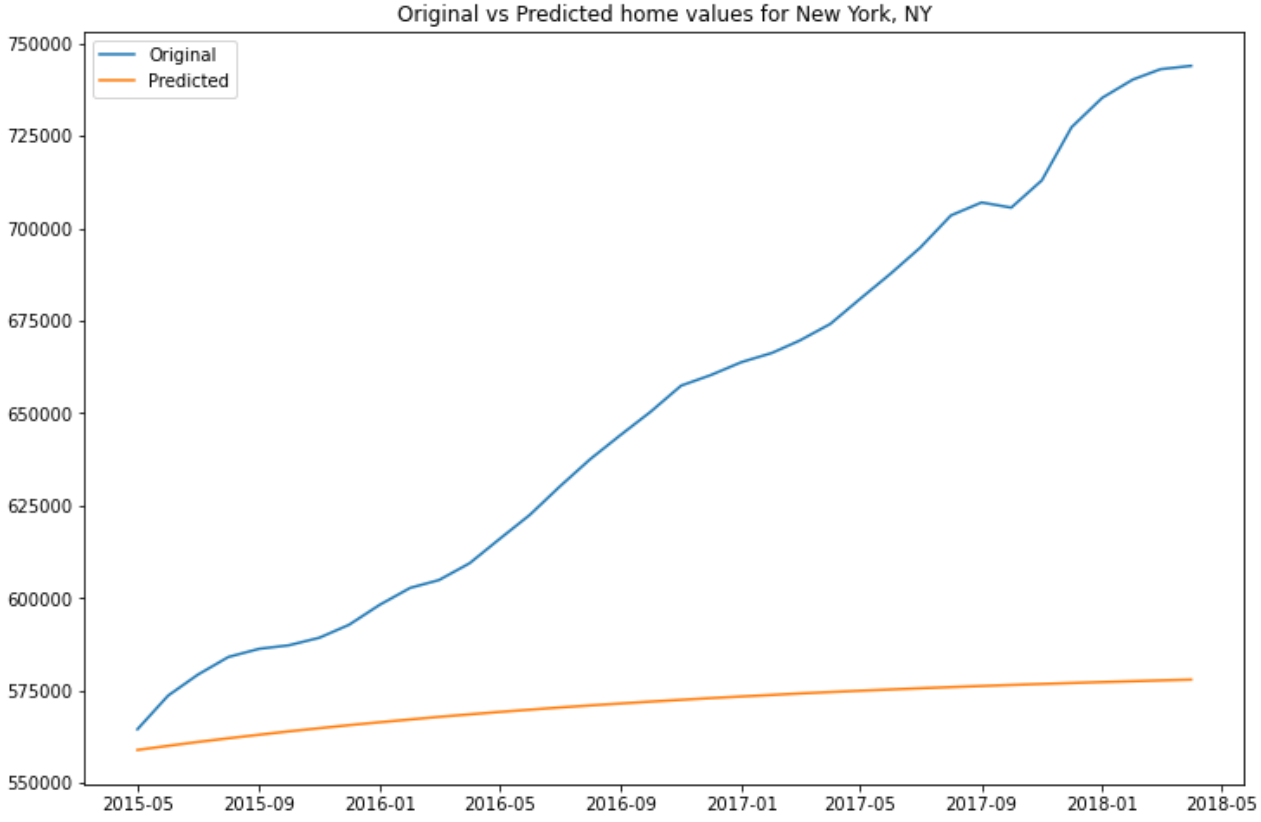
```
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.6985      0.258      2.708      0.007       0.193       1.204
ma.L1         -0.6732      0.263     -2.562      0.010      -1.188      -0.158
sigma2      1.208e+06   1.82e+04     66.240      0.000    1.17e+06    1.24e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   28.76   Jarque-Bera (JB):              4354
8.18
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               11.04   Skew:                            -
5.78
Prob(H) (two-sided):                  0.00   Kurtosis:                         6
9.71
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 53964.580442025406
                               SARIMAX Results
==============================================================================
Dep. Variable:        Los Angeles, CA   No. Observations:                  229
Model:                 ARIMA(1, 1, 1)   Log Likelihood               -2217.648
Date:                Fri, 13 May 2022   AIC                           4441.295
Time:                        12:02:38   BIC                           4451.583
Sample:                    04-01-1996   HQIC                          4445.446
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9296      0.018     51.030      0.000       0.894       0.965
ma.L1         -0.8611      0.024    -35.795      0.000      -0.908      -0.814
sigma2      1.603e+07   1.08e-10   1.49e+17      0.000     1.6e+07     1.6e+07
==============================================================================
===
Ljung-Box (L1) (Q):                  146.96   Jarque-Bera (JB):                3
5.91
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.25   Skew:                            -
0.46
Prob(H) (two-sided):                  0.33   Kurtosis:
4.71
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 7.74e+
32. Standard errors may be unstable.
------------------------
------------------------
RMSE: 78171.42845160789
                               SARIMAX Results
==============================================================================
Dep. Variable:           San Jose, CA   No. Observations:                  229
```

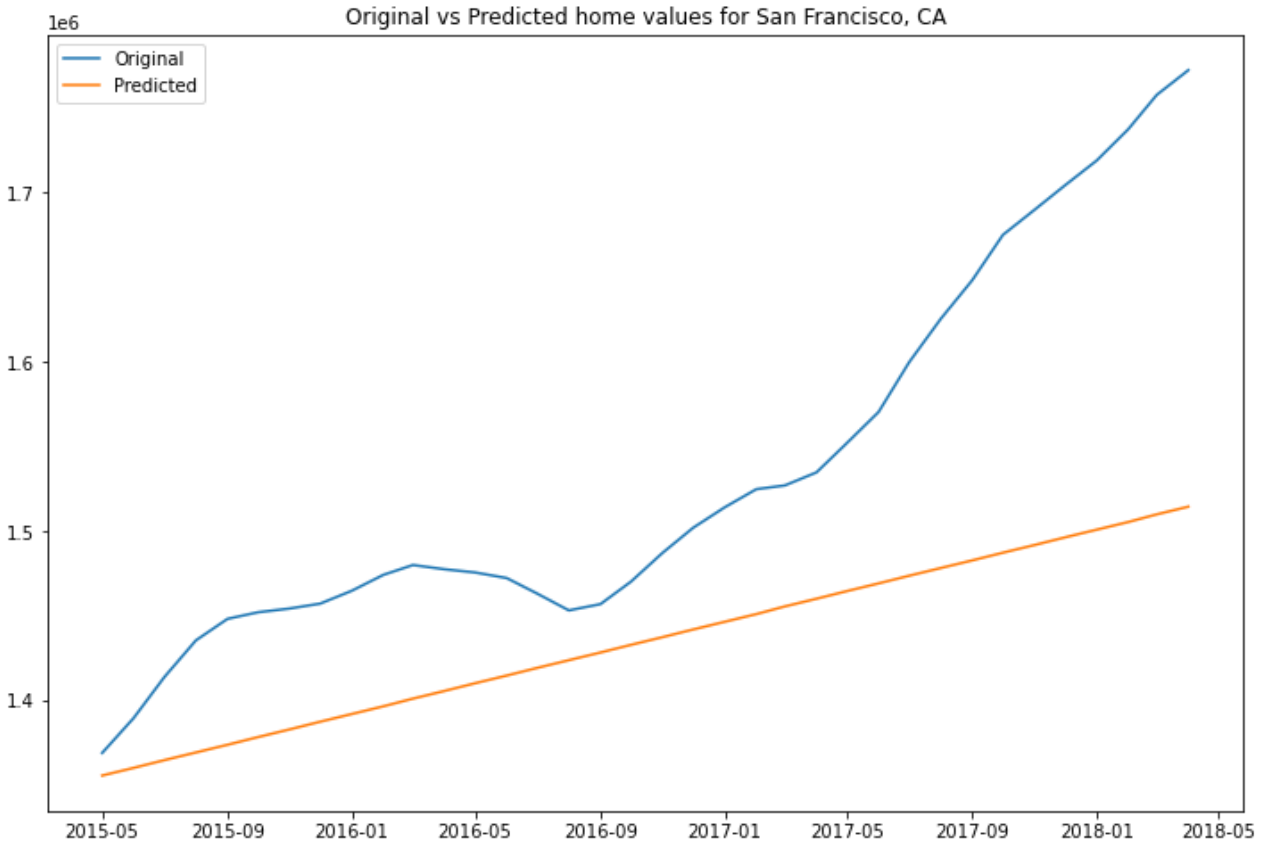
```
Model:                 ARIMA(1, 1, 1)   Log Likelihood               -2247.192
Date:                Fri, 13 May 2022   AIC                         4500.384
Time:                        12:02:39   BIC                         4510.672
Sample:                    04-01-1996   HQIC                        4504.535
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9308      0.022     41.557      0.000       0.887       0.975
ma.L1         -0.8776      0.028    -31.836      0.000      -0.932      -0.824
sigma2      2.095e+07   7.37e-11   2.84e+17      0.000     2.1e+07     2.1e+07
===================================================================================
Ljung-Box (L1) (Q):                 180.13   Jarque-Bera (JB):
5.82
Prob(Q):                              0.00   Prob(JB):
0.05
Heteroskedasticity (H):               1.41   Skew:                             -
0.25
Prob(H) (two-sided):                  0.13   Kurtosis:
3.61
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 1.43e+
33. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 141264.17840082556
                               SARIMAX Results
==============================================================================
Dep. Variable:            Chicago, IL   No. Observations:                  229
Model:                 ARIMA(1, 1, 1)   Log Likelihood               -2070.628
Date:                Fri, 13 May 2022   AIC                         4147.255
Time:                        12:02:39   BIC                         4157.543
Sample:                    04-01-1996   HQIC                        4151.406
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9594      0.013     72.413      0.000       0.933       0.985
ma.L1         -0.9316      0.016    -57.706      0.000      -0.963      -0.900
sigma2      4.311e+06   9.09e-10   4.74e+15      0.000    4.31e+06    4.31e+06
===================================================================================
Ljung-Box (L1) (Q):                 186.42   Jarque-Bera (JB):                 5
0.85
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.83   Skew:                             -
0.01
Prob(H) (two-sided):                  0.01   Kurtosis:
5.31
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 2.57e+
```

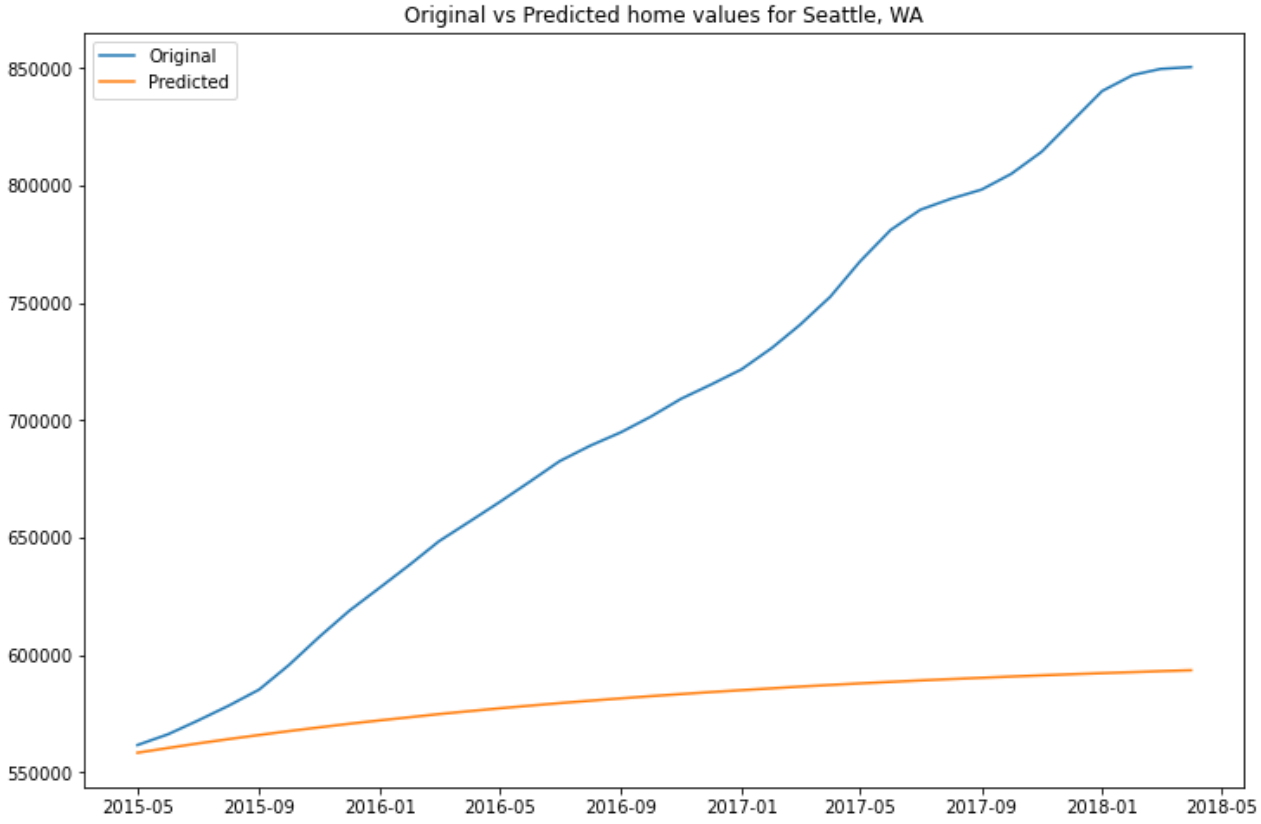
```
30. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 22740.16271525774
                               SARIMAX Results
==========================================================================================
Dep. Variable:          Baltimore, MD   No. Observations:                  229
Model:                 ARIMA(1, 1, 1)   Log Likelihood               -1884.281
Date:                Fri, 13 May 2022   AIC                           3774.562
Time:                        12:02:39   BIC                           3784.850
Sample:                    04-01-1996   HQIC                          3778.712
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9210      0.014     65.891      0.000       0.894       0.948
ma.L1         -0.8099      0.021    -38.766      0.000      -0.851      -0.769
sigma2      6.452e+05   4.28e+04     15.066      0.000    5.61e+05    7.29e+05
==============================================================================
===
Ljung-Box (L1) (Q):                 132.83   Jarque-Bera (JB):                 7
9.60
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               2.29   Skew:                            -
0.72
Prob(H) (two-sided):                  0.00   Kurtosis:
5.51
==========================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
-----------------------
-----------------------
RMSE: 9908.089549065977
                               SARIMAX Results
==========================================================================================
Dep. Variable:              Boston, MA   No. Observations:                  229
Model:                 ARIMA(1, 1, 1)   Log Likelihood               -2077.734
Date:                Fri, 13 May 2022   AIC                           4161.468
Time:                        12:02:40   BIC                           4171.756
Sample:                    04-01-1996   HQIC                          4165.619
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9697      0.013     75.145      0.000       0.944       0.995
ma.L1         -0.9362      0.018    -52.073      0.000      -0.971      -0.901
sigma2      4.628e+06    4.7e-11   9.86e+16      0.000    4.63e+06    4.63e+06
==============================================================================
===
Ljung-Box (L1) (Q):                 114.74   Jarque-Bera (JB):                13
8.57
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               2.31   Skew:
0.52
Prob(H) (two-sided):                  0.00   Kurtosis:
6.68
==========================================================================================
===
```

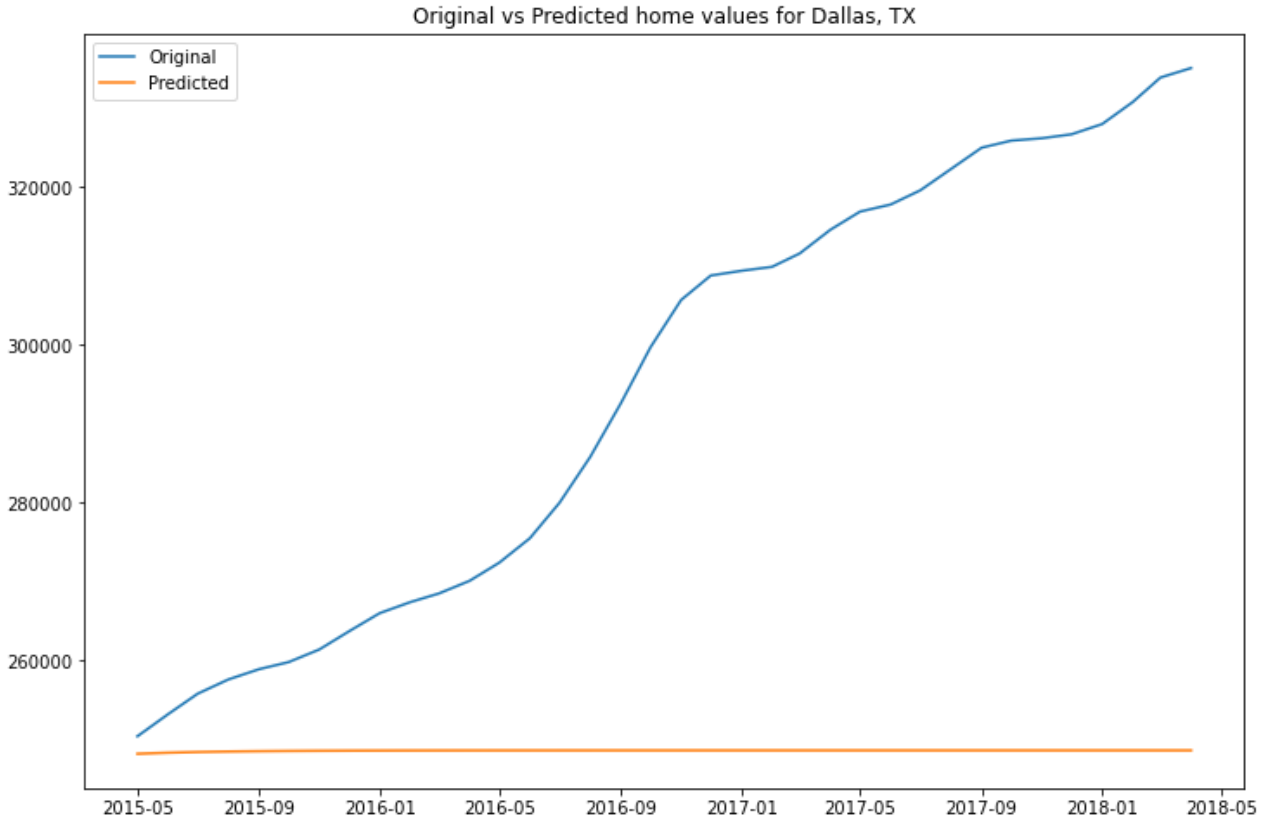
```
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 3.65e+
33. Standard errors may be unstable.
------------------------
------------------------
RMSE: 75745.33148262749
```

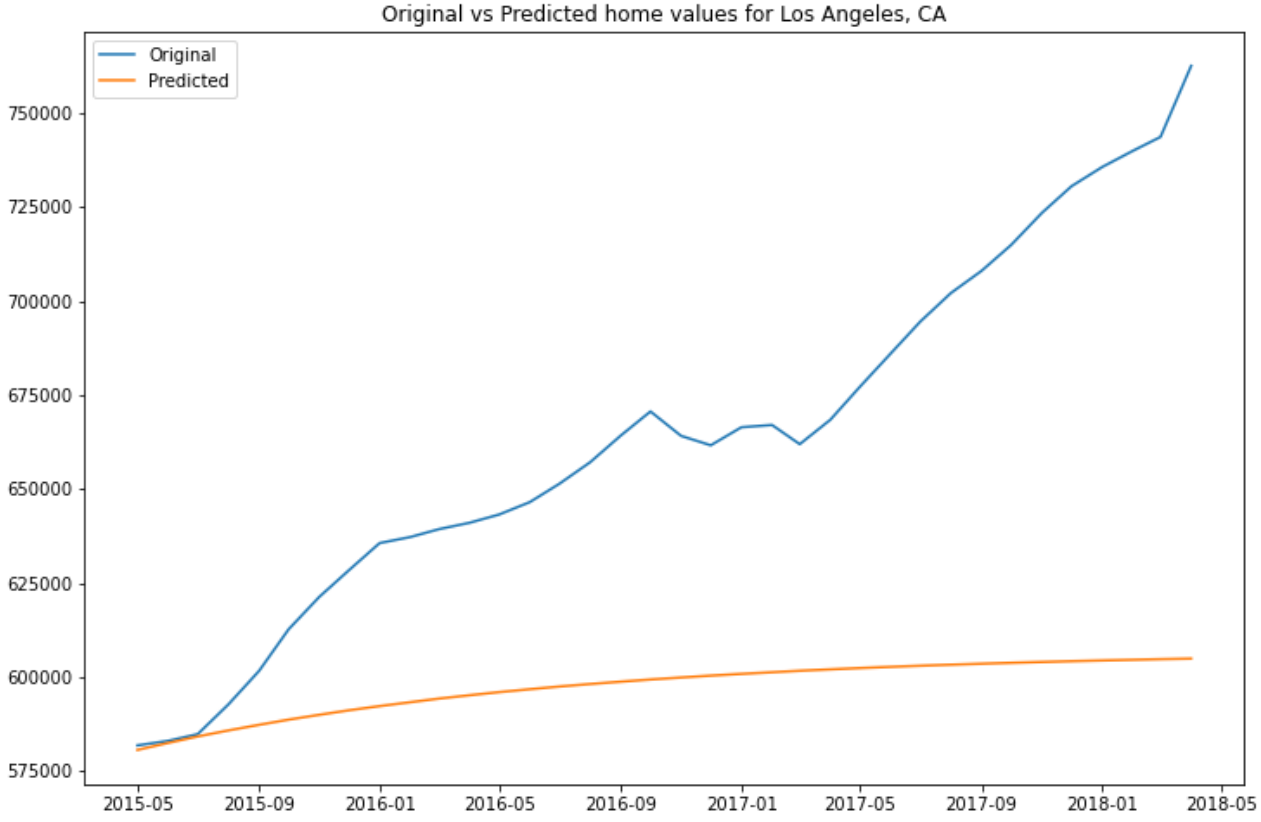


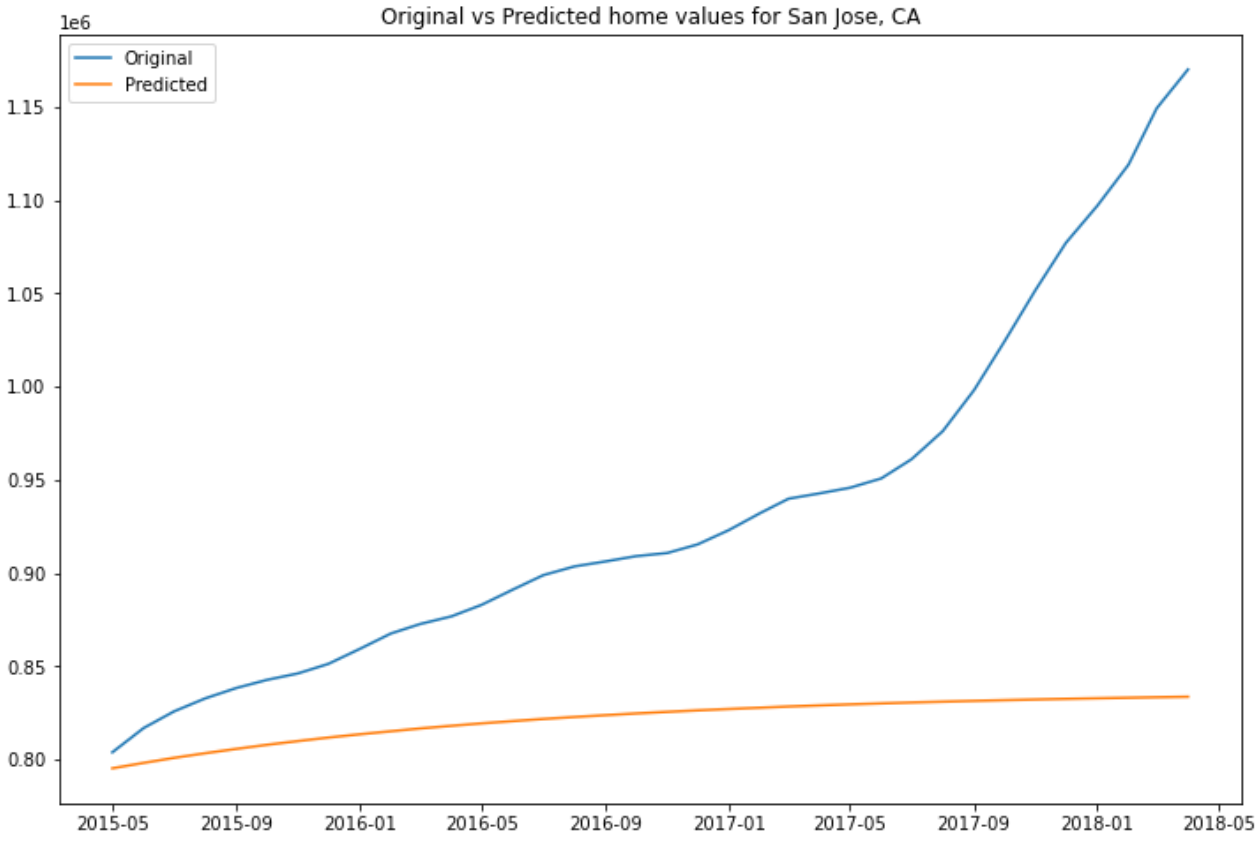


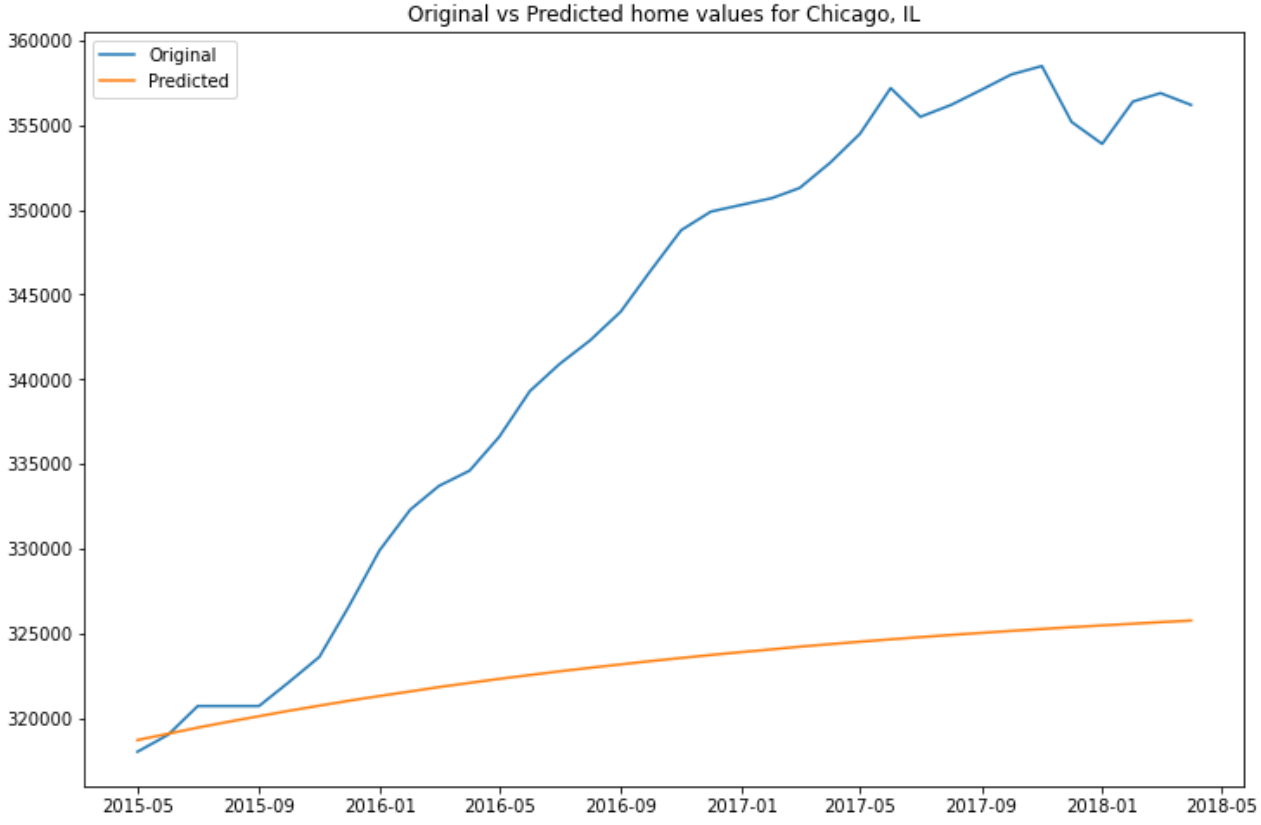

Original vs Predicted home values for San Francisco, CA
1e6
Original
Predicted
1.7
1.6
1.5
1.4
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

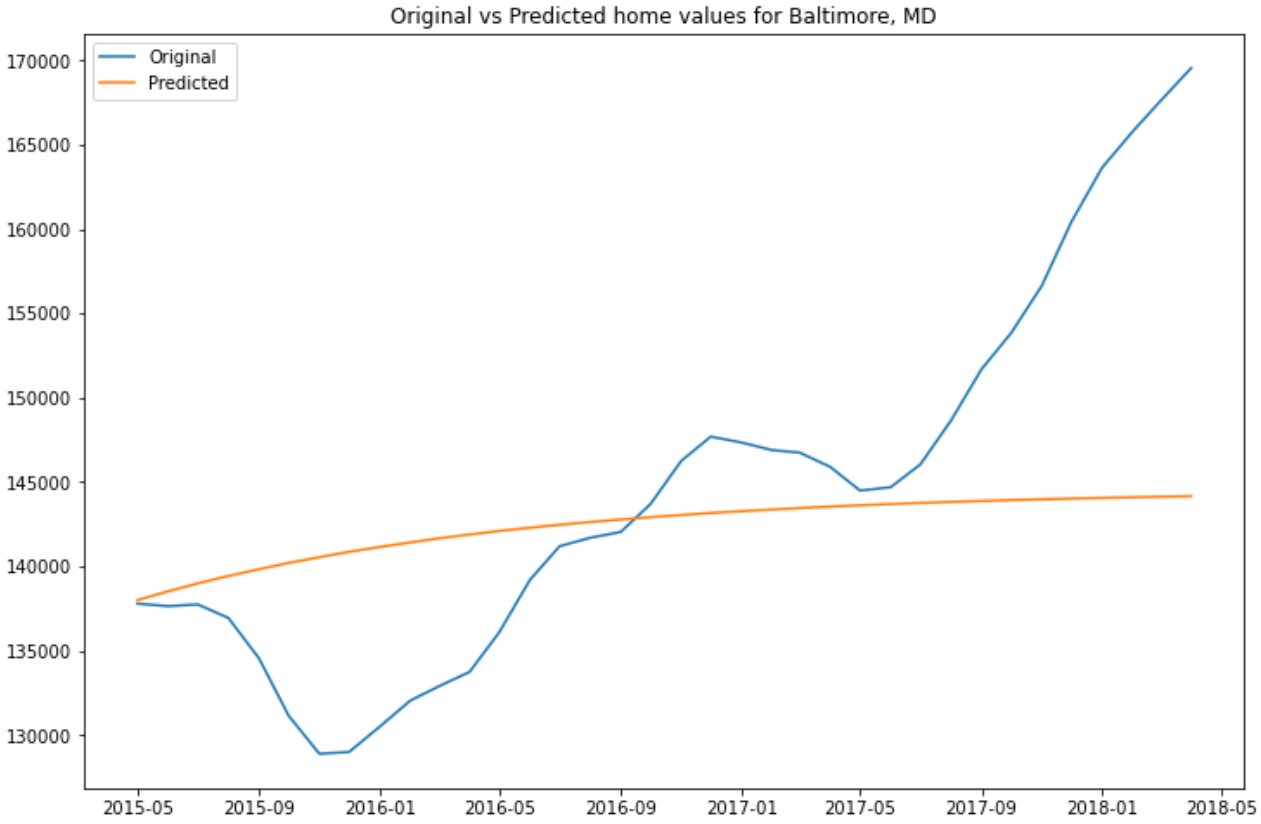

Original vs Predicted home values for Seattle, WA
Original
Predicted
850000
800000
750000
700000
650000
600000
550000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

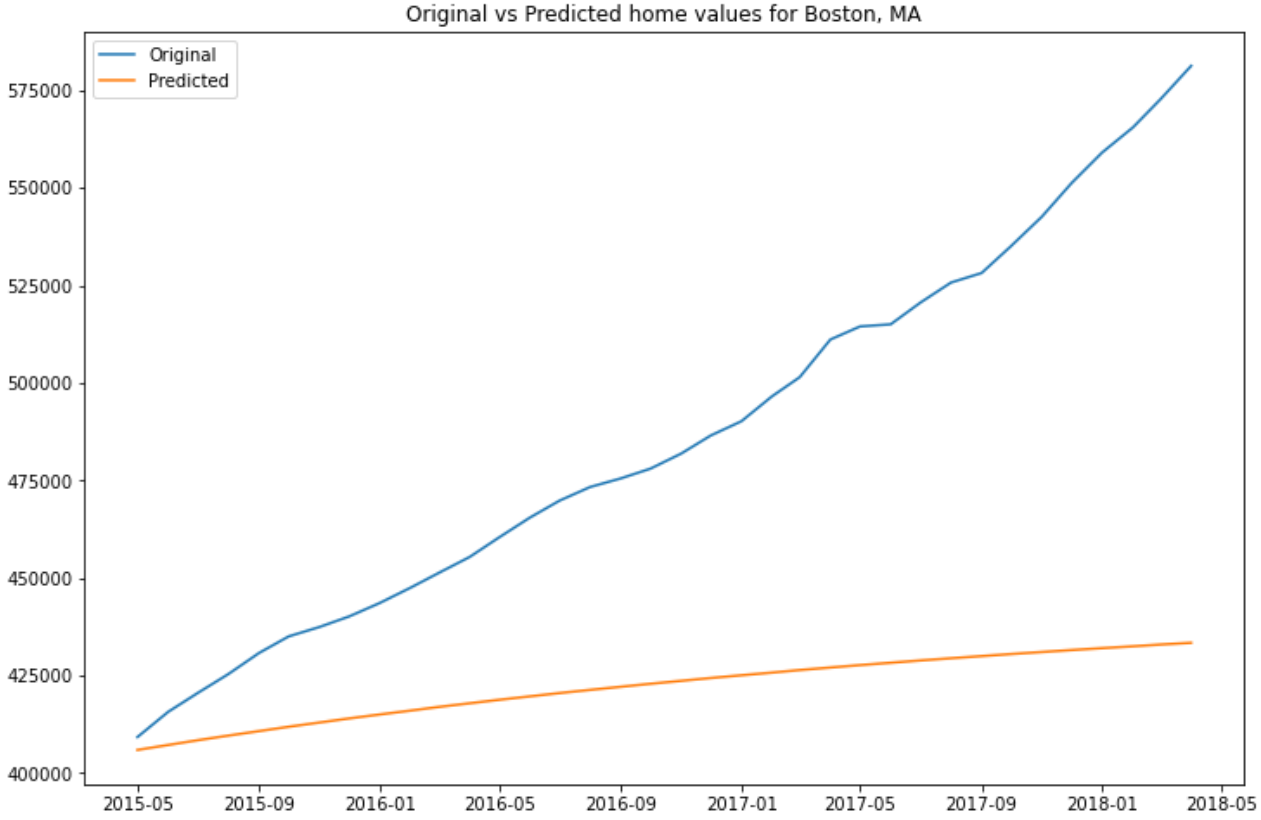

Original vs Predicted home values for Dallas, TX
Original
Predicted
320000
300000
280000
260000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Los Angeles, CA
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for San Jose, CA
1e6
Original
Predicted
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Chicago, IL
Original
Predicted
360000
355000
350000
345000
340000
335000
330000
325000
320000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05





In [70]:
```
# mean rmse for every city
np.mean(rmse_list)
```

Out[70]:
```
80278.61371440972
```

In [71]:
```
# still pretty rough, try increasing the differencing
rmse_list = []
```

```
for city in city_list:
    city_model = arima_mod(city)
    city_model.model(train, test, 1,2,1)
    city_model.plot(test)
    rmse_list.append(city_model.rmse_)
```

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                Washington, DC   No. Observations:                  229
Model:                 ARIMA(1, 2, 1)   Log Likelihood               -1863.778
Date:                Fri, 13 May 2022   AIC                           3733.555
Time:                        12:02:43   BIC                           3743.830
Sample:                    04-01-1996   HQIC                          3737.701
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1994      0.432      0.462      0.644      -0.646       1.045
ma.L1         -0.1372      0.423     -0.324      0.746      -0.967       0.693
sigma2      5.622e+05   2.49e+04     22.608      0.000    5.13e+05    6.11e+05
==============================================================================
===
Ljung-Box (L1) (Q):                  46.94   Jarque-Bera (JB):                10
2.28
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              15.97   Skew:
0.04
Prob(H) (two-sided):                  0.00   Kurtosis:
6.29
==========================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 48241.3990031015
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:978: UserWarning:

Non-invertible starting MA parameters found. Using zeros as starting parameters.

                               SARIMAX Results
==========================================================================================
Dep. Variable:                New York, NY   No. Observations:                  229
Model:                 ARIMA(1, 2, 1)   Log Likelihood               -2126.436
Date:                Fri, 13 May 2022   AIC                           4258.872
Time:                        12:02:43   BIC                           4269.147
Sample:                    04-01-1996   HQIC                          4263.018
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.8473      0.027     30.960      0.000       0.794       0.901
ma.L1         -0.9853      0.014    -71.397      0.000      -1.012      -0.958
sigma2      8.086e+06   1.06e-09   7.61e+15      0.000    8.09e+06    8.09e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   1.69   Jarque-Bera (JB):                31
1.15
```

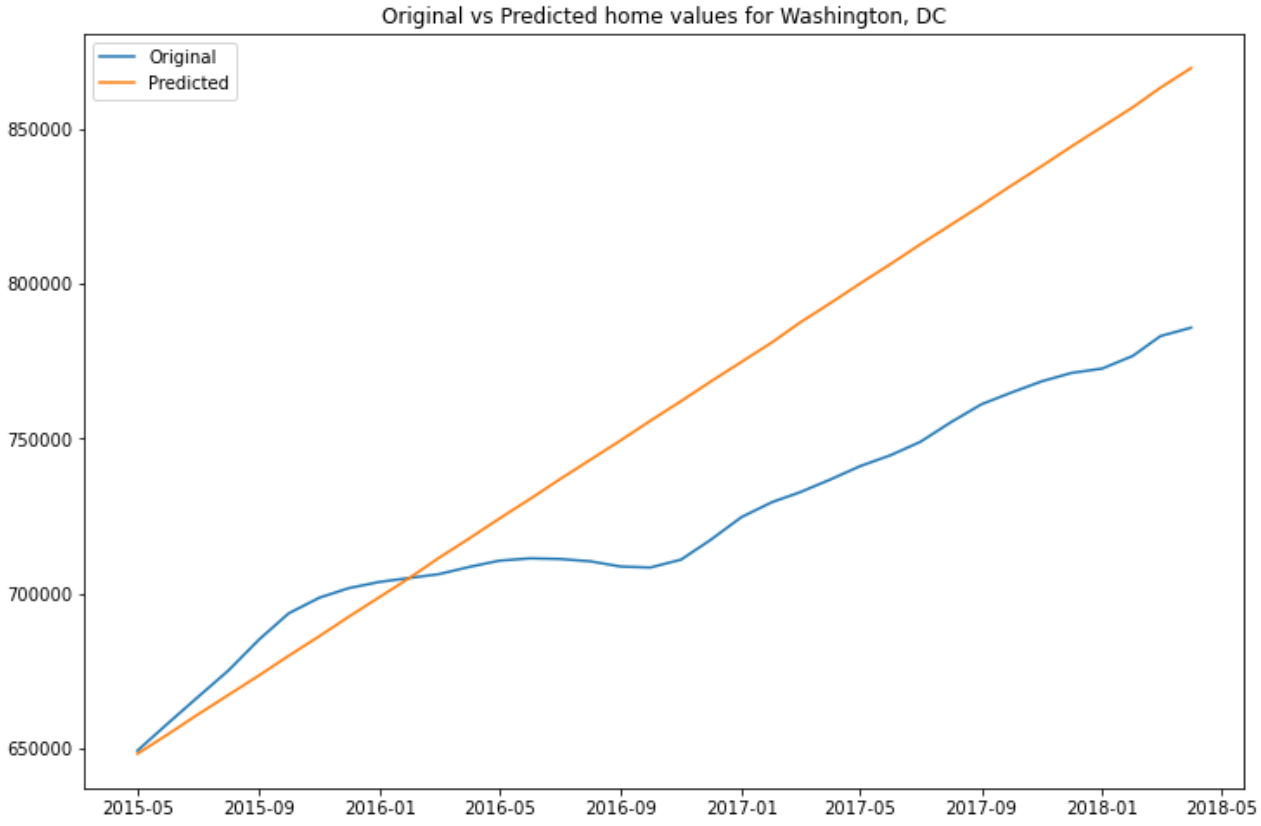
```
Prob(Q):                              0.19   Prob(JB):
0.00
Heteroskedasticity (H):              17.86   Skew:                             -
0.24
Prob(H) (two-sided):                  0.00   Kurtosis:
8.72
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 2.87e+
30. Standard errors may be unstable.
------------------------
------------------------
RMSE: 64937.83688840026
                           SARIMAX Results
================================================================================
Dep. Variable:      San Francisco, CA   No. Observations:                  229
Model:                 ARIMA(1, 2, 1)   Log Likelihood               -2145.107
Date:                Fri, 13 May 2022   AIC                           4296.215
Time:                        12:02:44   BIC                           4306.490
Sample:                  04-01-1996   HQIC                          4300.361
                       - 04-01-2015
Covariance Type:                 opg
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
ar.L1         -0.3551      3.717     -0.096      0.924      -7.641       6.931
ma.L1          0.3612      3.703      0.098      0.922      -6.896       7.618
sigma2       8.22e+06   1.42e-05   5.79e+11      0.000    8.22e+06    8.22e+06
================================================================================
===
Ljung-Box (L1) (Q):                  20.08   Jarque-Bera (JB):                10
3.38
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              23.72   Skew:
0.17
Prob(H) (two-sided):                  0.00   Kurtosis:
6.29
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 2.44e+
28. Standard errors may be unstable.
------------------------
------------------------
RMSE: 160177.391307992
                           SARIMAX Results
================================================================================
Dep. Variable:           Seattle, WA   No. Observations:                  229
Model:                 ARIMA(1, 2, 1)   Log Likelihood               -1878.881
Date:                Fri, 13 May 2022   AIC                           3763.763
Time:                        12:02:45   BIC                           3774.037
Sample:                  04-01-1996   HQIC                          3767.909
                       - 04-01-2015
Covariance Type:                 opg
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
```

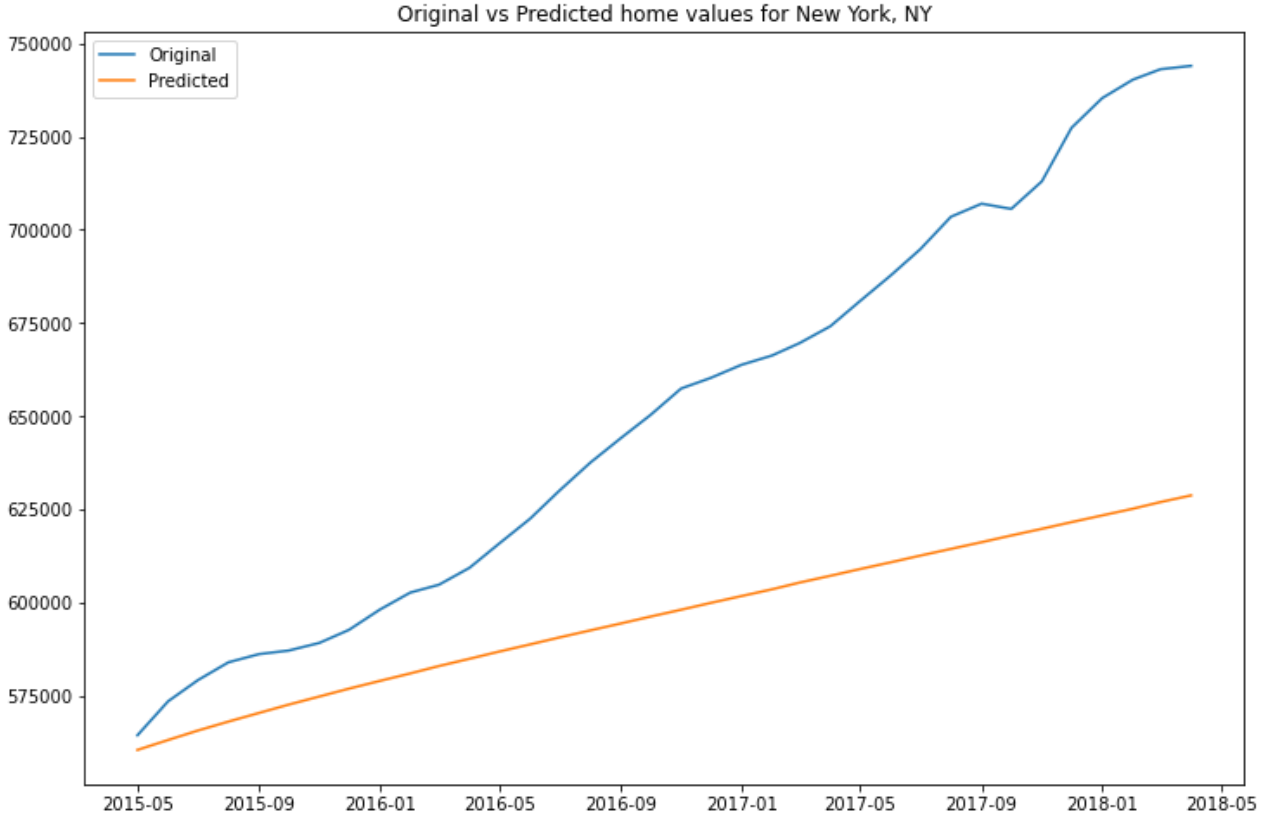
```
ar.L1         -0.1415      3.174     -0.045      0.964      -6.362       6.079
ma.L1          0.1540      3.158      0.049      0.961      -6.036       6.344
sigma2      9.005e+05   6.42e+04     14.029      0.000    7.75e+05    1.03e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   23.52   Jarque-Bera (JB):                 2
7.78
Prob(Q):                               0.00   Prob(JB):
0.00
Heteroskedasticity (H):               30.86   Skew:
0.27
Prob(H) (two-sided):                   0.00   Kurtosis:
4.63
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 45187.178478034446
                               SARIMAX Results
==============================================================================
Dep. Variable:             Dallas, TX   No. Observations:                  229
Model:                 ARIMA(1, 2, 1)   Log Likelihood               -1951.926
Date:                Fri, 13 May 2022   AIC                           3909.851
Time:                        12:02:45   BIC                           3920.126
Sample:                    04-01-1996   HQIC                          3913.997
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7374      0.442      1.667      0.095      -0.129       1.604
ma.L1         -0.7702      0.446     -1.728      0.084      -1.644       0.104
sigma2      1.719e+06   2.96e+04     58.170      0.000    1.66e+06    1.78e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   38.71   Jarque-Bera (JB):             7667
5.53
Prob(Q):                               0.00   Prob(JB):
0.00
Heteroskedasticity (H):               42.99   Skew:                            -
1.50
Prob(H) (two-sided):                   0.00   Kurtosis:                         9
2.99
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 5143.744347784151
                               SARIMAX Results
==============================================================================
Dep. Variable:        Los Angeles, CA   No. Observations:                  229
Model:                 ARIMA(1, 2, 1)   Log Likelihood               -2075.589
Date:                Fri, 13 May 2022   AIC                           4157.179
Time:                        12:02:46   BIC                           4167.454
Sample:                    04-01-1996   HQIC                          4161.325
                         - 04-01-2015
Covariance Type:                  opg
```

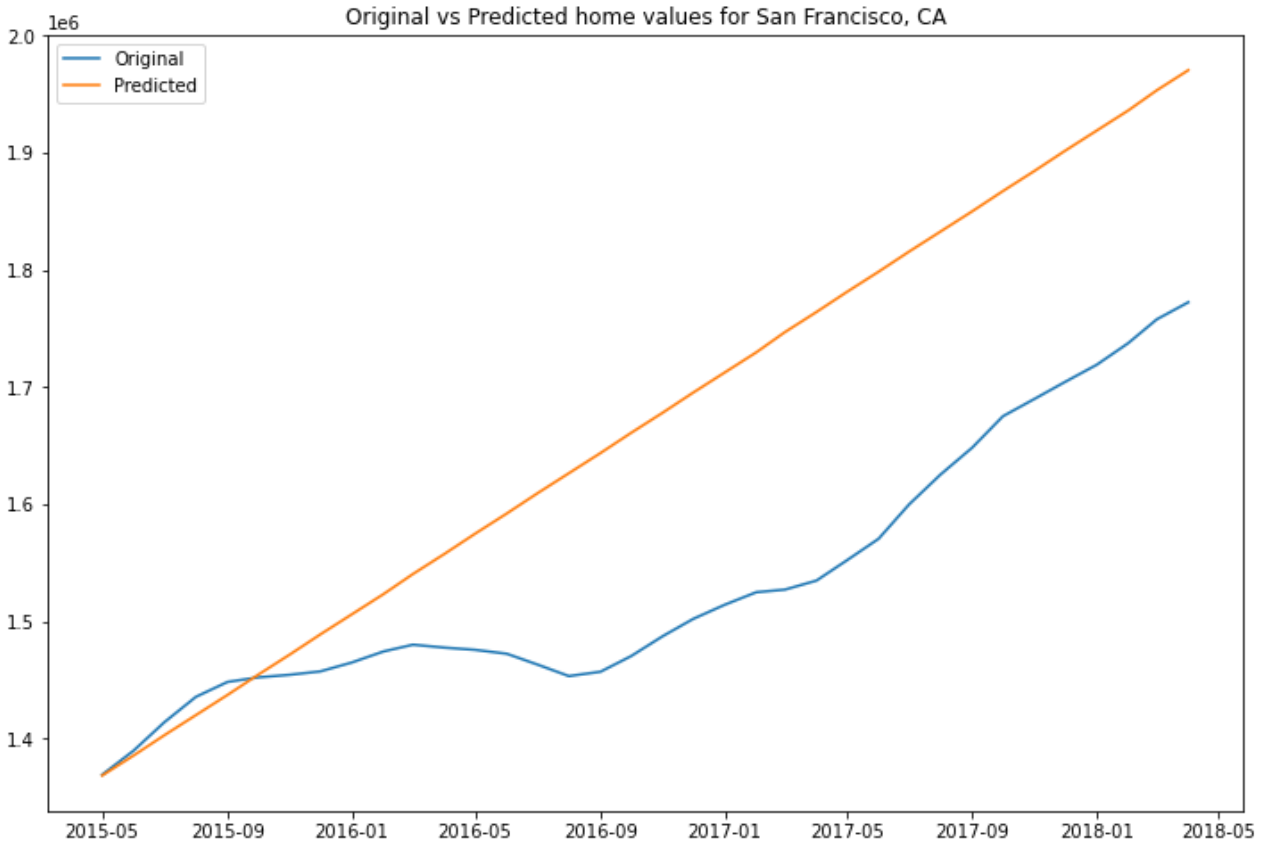
```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9301      0.033     27.852      0.000       0.865       0.996
ma.L1         -0.9890      0.019    -53.381      0.000      -1.025      -0.953
sigma2      5.153e+06   2.12e+05     24.267      0.000    4.74e+06    5.57e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   10.17   Jarque-Bera (JB):              1501
0.85
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               31.96   Skew:                            -
3.10
Prob(H) (two-sided):                  0.00   Kurtosis:                         4
2.35
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 31735.793768466134
                               SARIMAX Results
==============================================================================
Dep. Variable:           San Jose, CA   No. Observations:                  229
Model:                 ARIMA(1, 2, 1)   Log Likelihood               -2028.409
Date:                Fri, 13 May 2022   AIC                           4062.818
Time:                        12:02:46   BIC                           4073.093
Sample:                    04-01-1996   HQIC                          4066.964
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6197      4.860     -0.127      0.899     -10.146       8.906
ma.L1          0.6230      4.851      0.128      0.898      -8.886      10.132
sigma2      3.187e+06   1.21e+05     26.428      0.000    2.95e+06    3.42e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   2.86   Jarque-Bera (JB):                84
7.30
Prob(Q):                              0.09   Prob(JB):
0.00
Heteroskedasticity (H):               16.53   Skew:                            -
0.04
Prob(H) (two-sided):                  0.00   Kurtosis:                         1
2.46
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 33094.77626506625
                               SARIMAX Results
==============================================================================
Dep. Variable:             Chicago, IL   No. Observations:                  229
Model:                 ARIMA(1, 2, 1)   Log Likelihood               -1864.111
Date:                Fri, 13 May 2022   AIC                           3734.222
Time:                        12:02:46   BIC                           3744.497
```

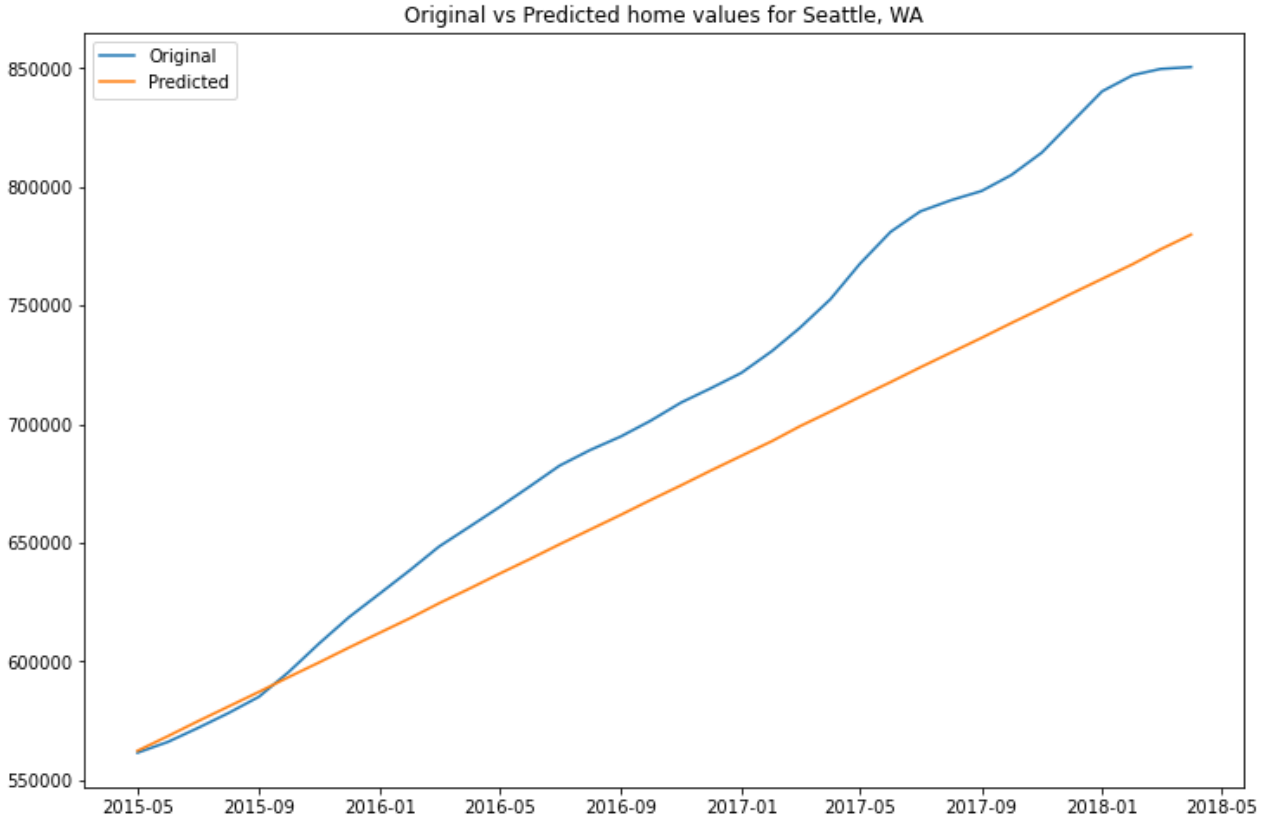
```
Sample:                     04-01-1996   HQIC                          3738.368
                          - 04-01-2015
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0752      1.533     -0.049      0.961      -3.079       2.929
ma.L1          0.0926      1.526      0.061      0.952      -2.899       3.084
sigma2      7.567e+05   2.98e+04     25.410      0.000    6.98e+05    8.15e+05
==============================================================================
===
Ljung-Box (L1) (Q):                   9.76   Jarque-Bera (JB):                77
2.43
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              31.38   Skew:
0.36
Prob(H) (two-sided):                  0.00   Kurtosis:                         1
2.01
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 4495.625351992376
                               SARIMAX Results
==============================================================================
Dep. Variable:          Baltimore, MD   No. Observations:                  229
Model:                 ARIMA(1, 2, 1)   Log Likelihood               -1751.737
Date:                Fri, 13 May 2022   AIC                           3509.475
Time:                        12:02:46   BIC                           3519.750
Sample:                     04-01-1996   HQIC                          3513.621
                          - 04-01-2015
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5638      0.214     -2.635      0.008      -0.983      -0.144
ma.L1          0.6515      0.208      3.136      0.002       0.244       1.059
sigma2      2.835e+05   1.31e+04     21.710      0.000    2.58e+05    3.09e+05
==============================================================================
===
Ljung-Box (L1) (Q):                   1.10   Jarque-Bera (JB):                57
2.05
Prob(Q):                              0.29   Prob(JB):
0.00
Heteroskedasticity (H):              22.29   Skew:
1.27
Prob(H) (two-sided):                  0.00   Kurtosis:                         1
0.35
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 19673.035432506487
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:966: UserWarning:
```

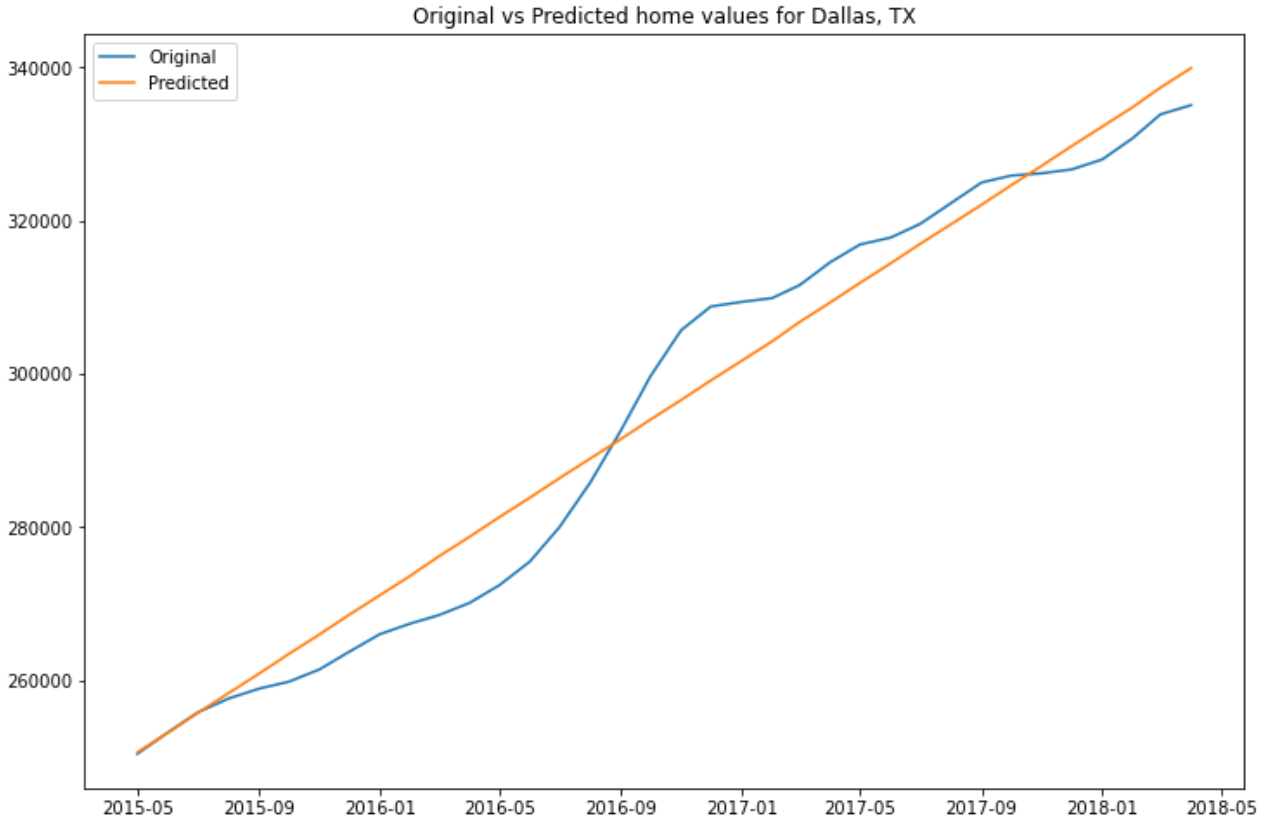
```
Non-stationary starting autoregressive parameters found. Using zeros as starting
parameters.

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:978: UserWarning:

Non-invertible starting MA parameters found. Using zeros as starting parameters.

                               SARIMAX Results
==========================================================================================
Dep. Variable:                      Boston, MA   No. Observations:                   229
Model:                  ARIMA(1, 2, 1)   Log Likelihood               -1993.712
Date:                 Fri, 13 May 2022   AIC                           3993.424
Time:                         12:02:47   BIC                           4003.698
Sample:                     04-01-1996   HQIC                          3997.570
                          - 04-01-2015
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.8550      0.078     10.893      0.000       0.701       1.009
ma.L1         -0.9447      0.064    -14.857      0.000      -1.069      -0.820
sigma2      2.507e+06   8.82e+04     28.433      0.000    2.33e+06    2.68e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   2.59   Jarque-Bera (JB):             1031
9.91
Prob(Q):                              0.11   Prob(JB):
0.00
Heteroskedasticity (H):              22.23   Skew:                            -
1.26
Prob(H) (two-sided):                  0.00   Kurtosis:                         3
5.94
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
-----------------------
-----------------------
RMSE: 54461.512745274114
```

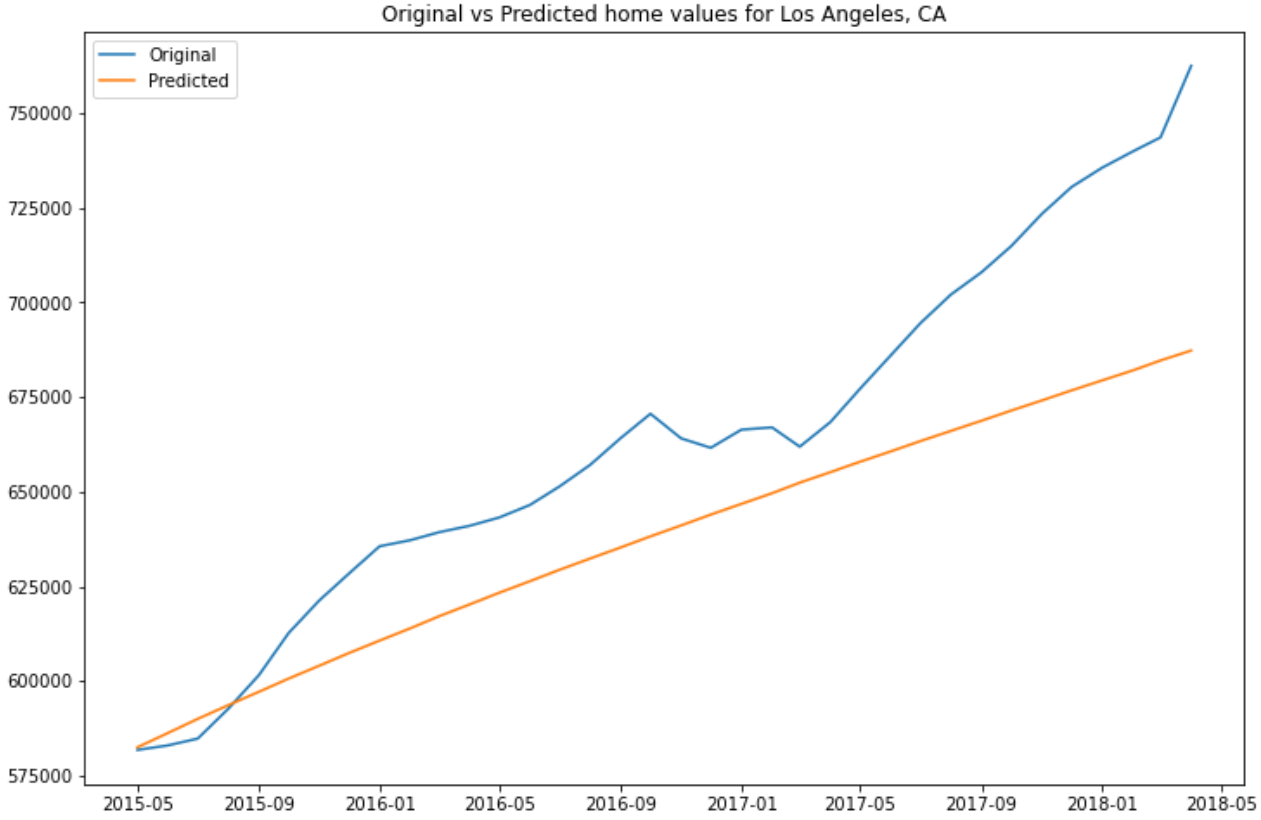

Original vs Predicted home values for Washington, DC
Original
Predicted
850000
800000
750000
700000
650000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

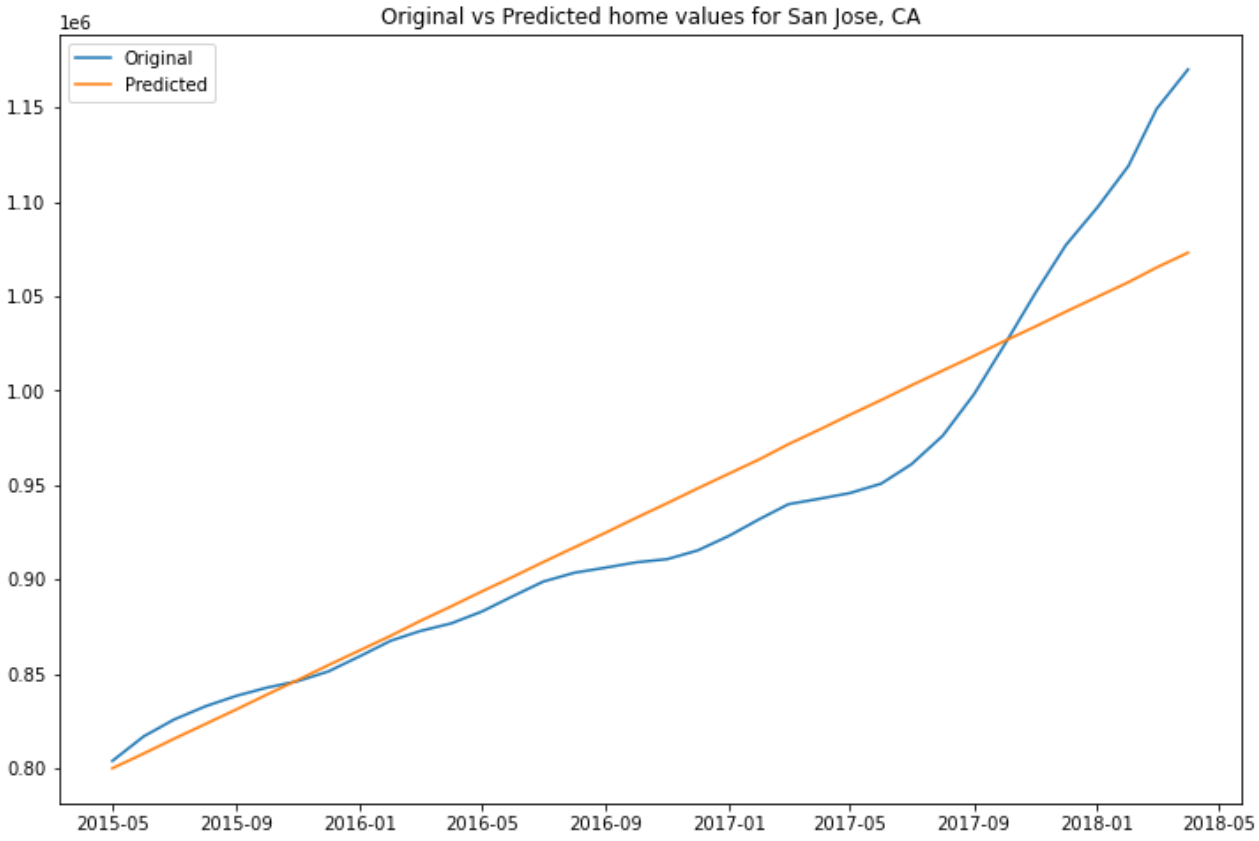

Original vs Predicted home values for New York, NY
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

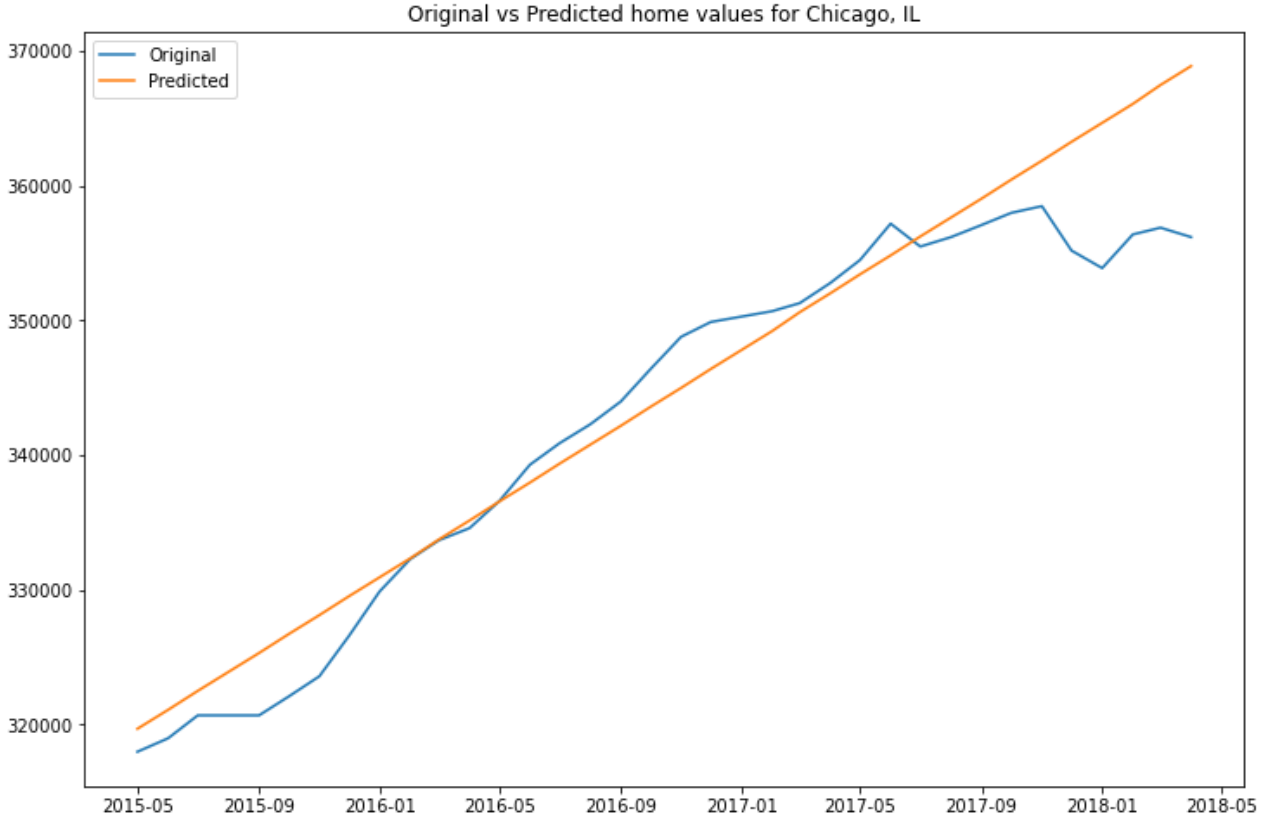

Original vs Predicted home values for San Francisco, CA
1e6
2.0
1.9
1.8
1.7
1.6
1.5
1.4
Original
Predicted
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

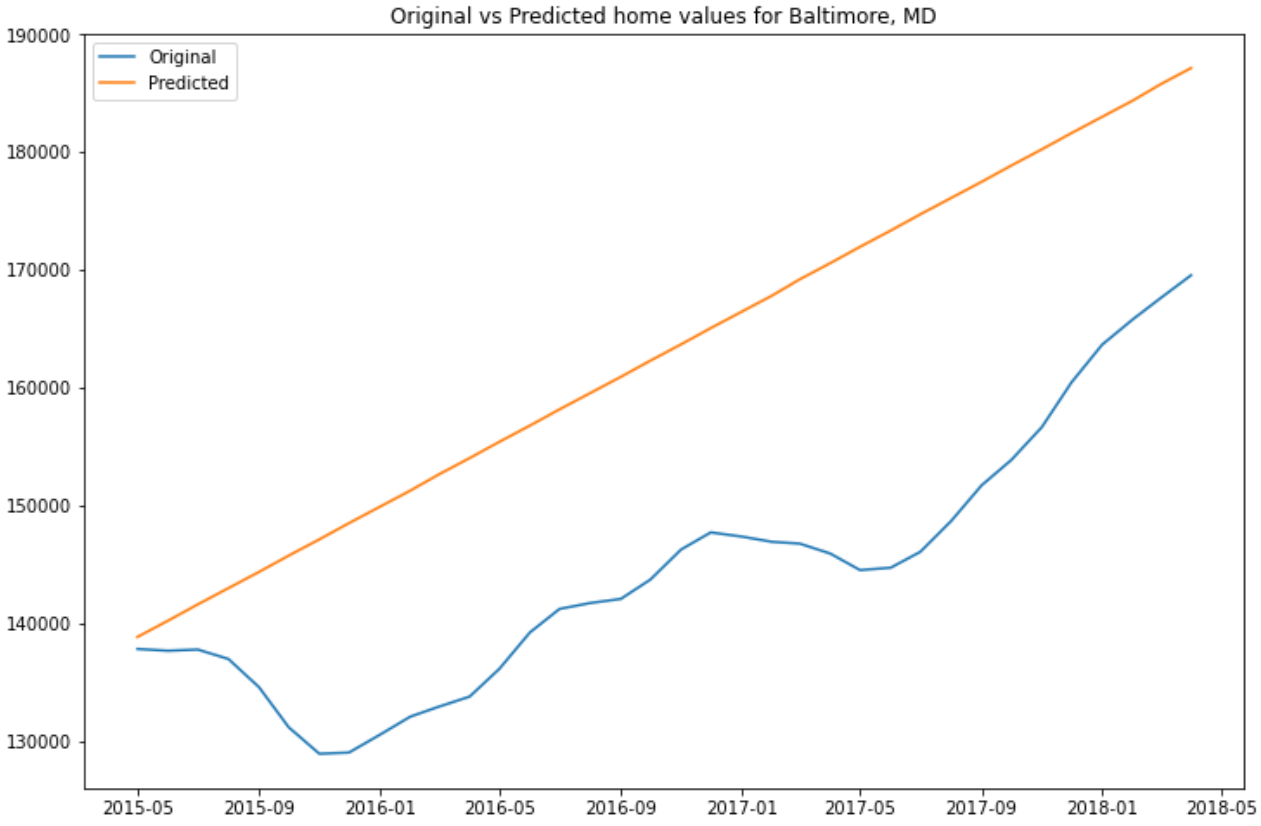

Original vs Predicted home values for Seattle, WA
850000
800000
750000
700000
650000
600000
550000
Original
Predicted
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

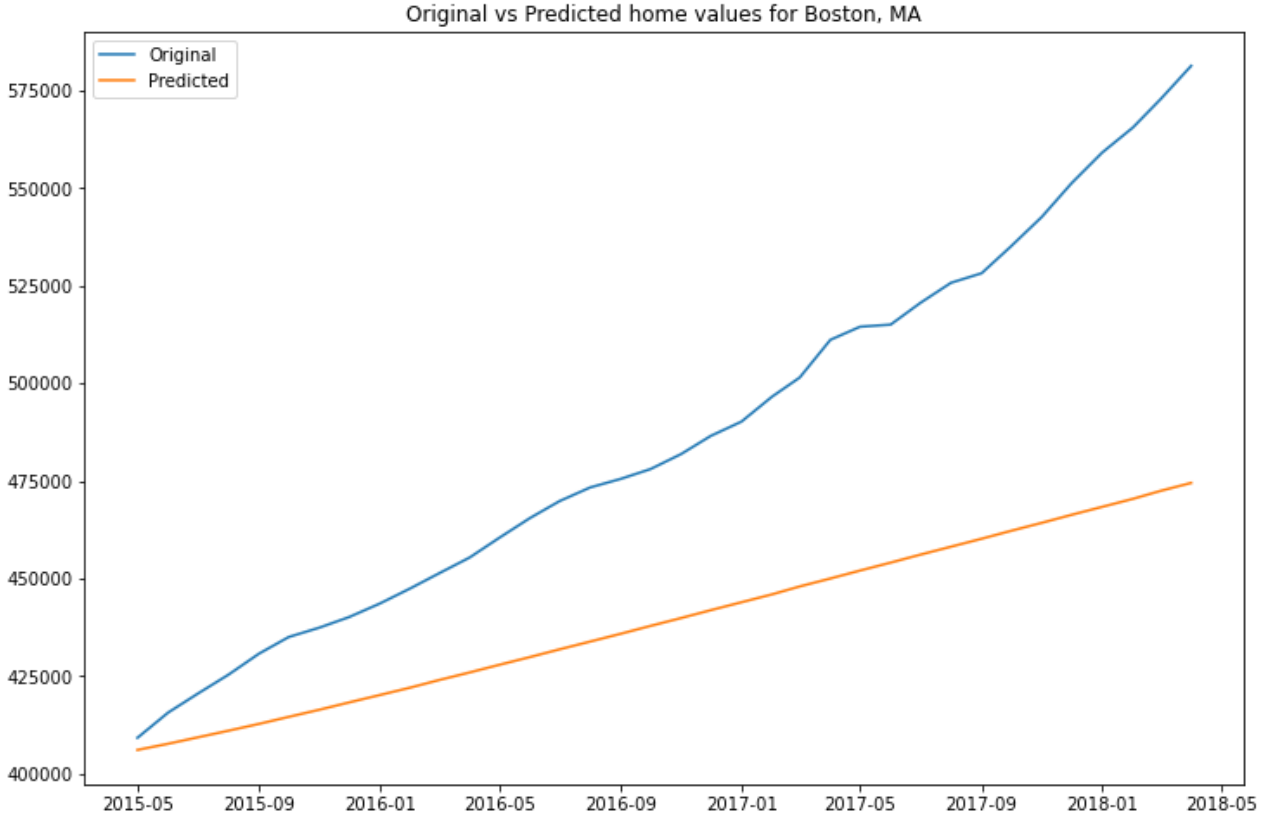

Original vs Predicted home values for Dallas, TX
Original
Predicted
340000
320000
300000
280000
260000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Los Angeles, CA
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for San Jose, CA
1e6
Original
Predicted
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Chicago, IL
Original
Predicted
370000
360000
350000
340000
330000
320000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05





In [72]:
```
# big improvement, still needs work
np.mean(rmse_list)
```

Out[72]: 46714.82935886177

In [73]:
```
# differencing helped, try out the moving average
rmse_list = []
```

```
for city in city_list:
    city_model = arima_mod(city)
    city_model.model(train, test, 1,2,2)
    city_model.plot(test)
    rmse_list.append(city_model.rmse_)
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:          Washington, DC   No. Observations:                  229
Model:                 ARIMA(1, 2, 2)   Log Likelihood               -1859.012
Date:                Fri, 13 May 2022   AIC                           3726.025
Time:                        12:02:51   BIC                           3739.725
Sample:                    04-01-1996   HQIC                          3731.553
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4563      2.210     -0.206      0.836      -4.788       3.875
ma.L1          0.4987      2.206      0.226      0.821      -3.825       4.822
ma.L2          0.0325      0.083      0.389      0.697      -0.131       0.196
sigma2      7.571e+05   4.68e+04     16.194      0.000    6.65e+05    8.49e+05
===================================================================================
===
Ljung-Box (L1) (Q):                  50.35   Jarque-Bera (JB):                10
1.72
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              20.22   Skew:
0.01
Prob(H) (two-sided):                  0.00   Kurtosis:
6.28
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 48366.372303319426
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:978: UserWarning:

Non-invertible starting MA parameters found. Using zeros as starting parameters.

                               SARIMAX Results
==============================================================================
Dep. Variable:          New York, NY   No. Observations:                  229
Model:                 ARIMA(1, 2, 2)   Log Likelihood               -2127.068
Date:                Fri, 13 May 2022   AIC                           4262.135
Time:                        12:02:52   BIC                           4275.835
Sample:                    04-01-1996   HQIC                          4267.663
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.8268      0.040     20.814      0.000       0.749       0.905
ma.L1         -0.9617      0.057    -16.747      0.000      -1.074      -0.849
ma.L2         -0.0212      0.057     -0.372      0.710      -0.133       0.091
sigma2      7.058e+06   2.68e-09   2.64e+15      0.000    7.06e+06    7.06e+06
===================================================================================
===
```

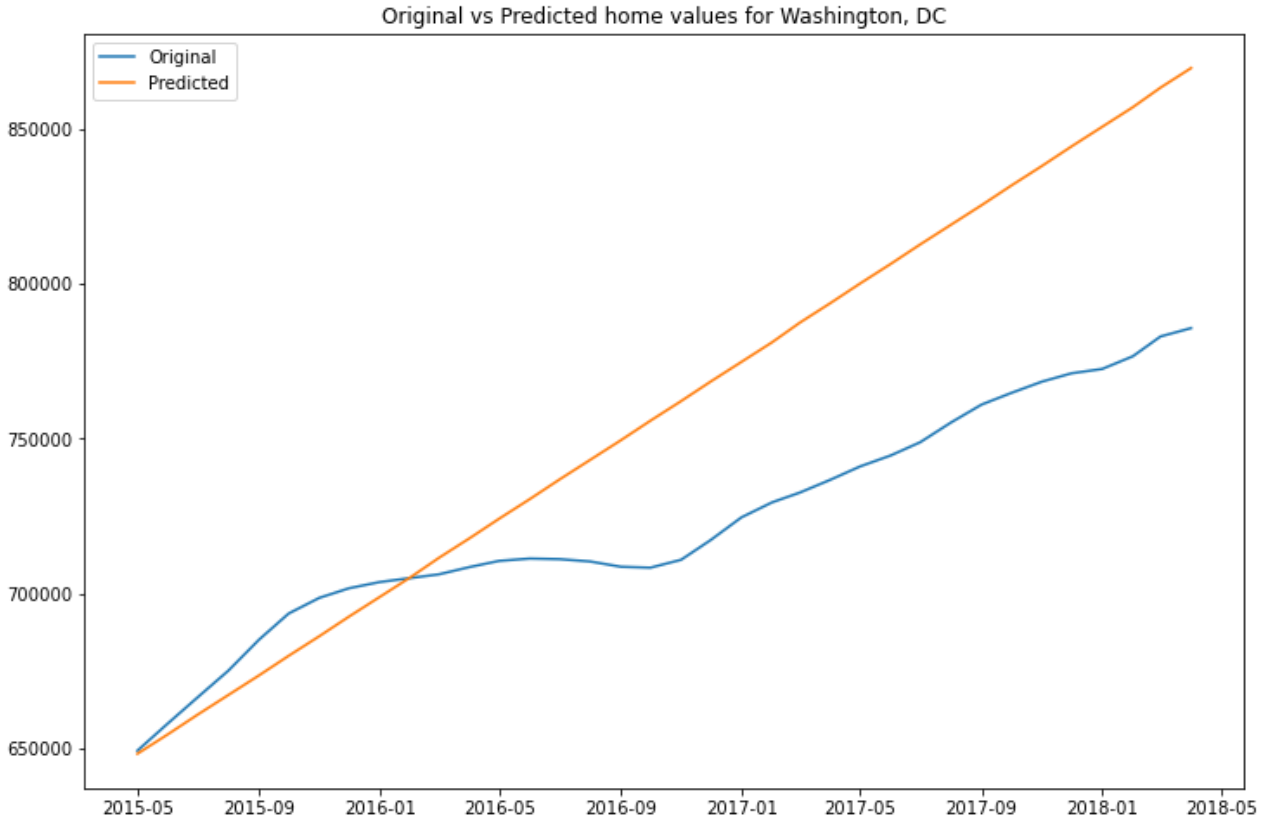
```
Ljung-Box (L1) (Q):                   1.96   Jarque-Bera (JB):               30
0.38
Prob(Q):                              0.16   Prob(JB):
0.00
Heteroskedasticity (H):              16.01   Skew:                            -
0.22
Prob(H) (two-sided):                  0.00   Kurtosis:
8.62
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 6.32e+
30. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 66378.57569196516
                               SARIMAX Results
================================================================================
Dep. Variable:      San Francisco, CA   No. Observations:                  229
Model:                 ARIMA(1, 2, 2)   Log Likelihood               -2141.449
Date:                Fri, 13 May 2022   AIC                           4290.899
Time:                        12:02:52   BIC                           4304.599
Sample:                  04-01-1996   HQIC                          4296.427
                       - 04-01-2015
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9413      0.025     37.092      0.000       0.892       0.991
ma.L1         -0.9846      0.070    -14.002      0.000      -1.122      -0.847
ma.L2         -0.0154      0.012     -1.276      0.202      -0.039       0.008
sigma2      9.126e+06    7.7e-09   1.19e+15      0.000    9.13e+06    9.13e+06
================================================================================
===
Ljung-Box (L1) (Q):                  21.47   Jarque-Bera (JB):                8
9.41
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              19.22   Skew:
0.39
Prob(H) (two-sided):                  0.00   Kurtosis:
5.98
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 2.77e+
29. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 68937.68288753173
                               SARIMAX Results
================================================================================
Dep. Variable:            Seattle, WA   No. Observations:                  229
Model:                 ARIMA(1, 2, 2)   Log Likelihood               -1878.129
Date:                Fri, 13 May 2022   AIC                           3764.258
Time:                        12:02:53   BIC                           3777.958
Sample:                  04-01-1996   HQIC                          3769.786
                       - 04-01-2015
Covariance Type:                 opg
```

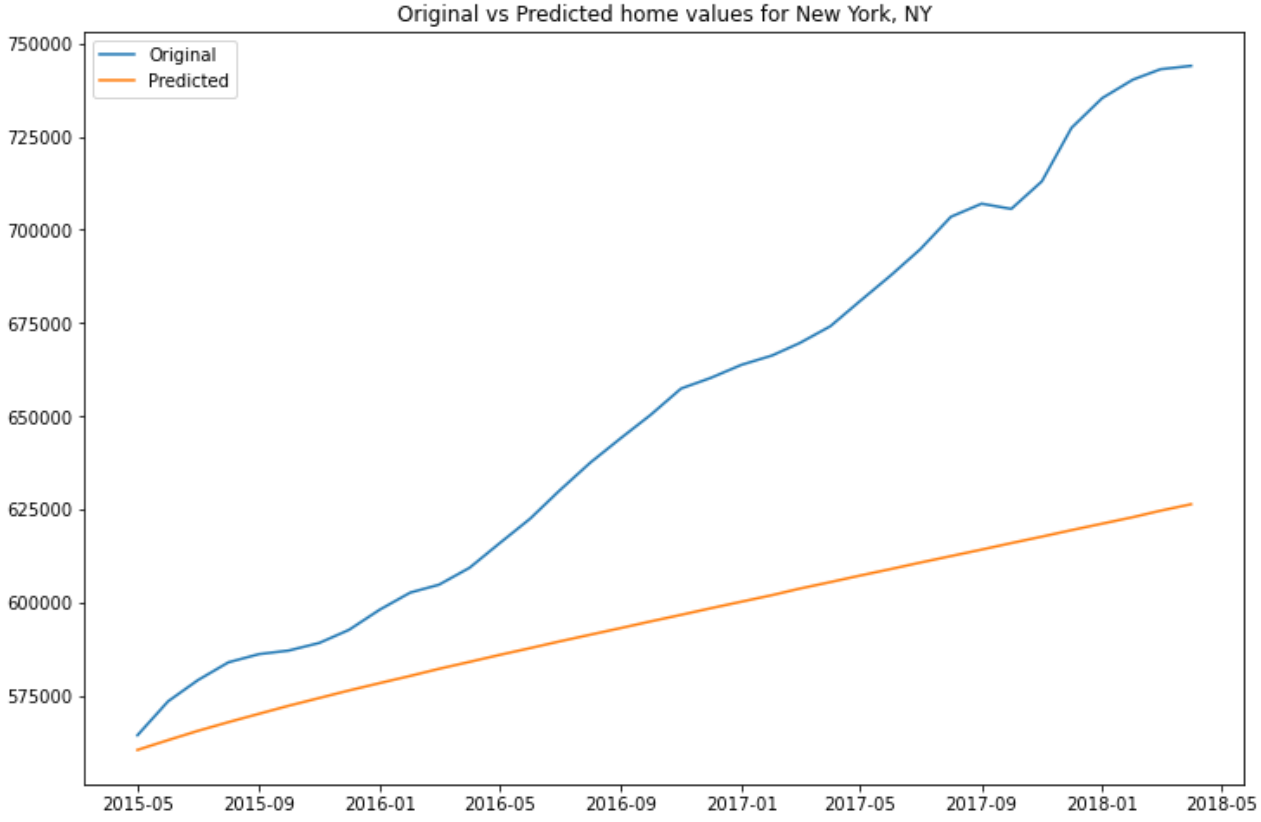
```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.8720      0.356      2.451      0.014       0.175       1.569
ma.L1         -0.8643      0.353     -2.450      0.014      -1.556      -0.173
ma.L2         -0.0204      0.017     -1.200      0.230      -0.054       0.013
sigma2      8.896e+05   6.55e+04     13.572      0.000    7.61e+05    1.02e+06
==============================================================================
===
Ljung-Box (L1) (Q):                  23.59   Jarque-Bera (JB):                 2
7.97
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              28.97   Skew:
0.33
Prob(H) (two-sided):                  0.00   Kurtosis:
4.59
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
-----------------------
-----------------------
RMSE: 47541.6581736747
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:966: UserWarning:

Non-stationary starting autoregressive parameters found. Using zeros as starting
parameters.

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:978: UserWarning:

Non-invertible starting MA parameters found. Using zeros as starting parameters.

                               SARIMAX Results
==============================================================================
Dep. Variable:             Dallas, TX   No. Observations:                  229
Model:                 ARIMA(1, 2, 2)   Log Likelihood               -1956.044
Date:                Fri, 13 May 2022   AIC                           3920.088
Time:                        12:02:53   BIC                           3933.788
Sample:                    04-01-1996   HQIC                          3925.616
                         - 04-01-2015
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7799      0.782      0.998      0.318      -0.752       2.312
ma.L1         -0.8162      0.784     -1.042      0.298      -2.352       0.720
ma.L2          0.0085      0.030      0.284      0.776      -0.050       0.067
sigma2      1.323e+06   1.76e+04     75.180      0.000    1.29e+06    1.36e+06
==============================================================================
===
Ljung-Box (L1) (Q):                  38.12   Jarque-Bera (JB):              7608
1.17
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              40.44   Skew:                            -
1.55
Prob(H) (two-sided):                  0.00   Kurtosis:                         9
2.63
==============================================================================
===
```

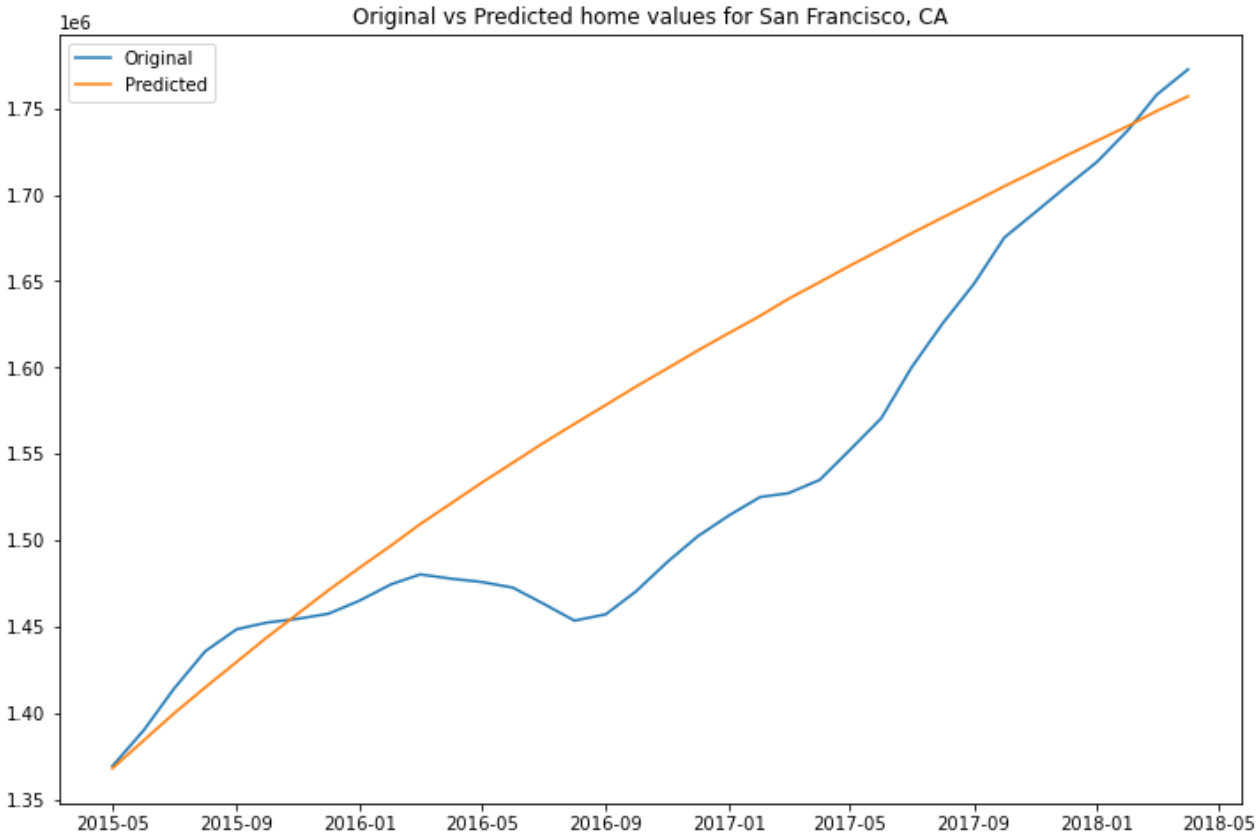
```
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 5148.832363111618
                               SARIMAX Results
==============================================================================
Dep. Variable:        Los Angeles, CA   No. Observations:                  229
Model:                 ARIMA(1, 2, 2)   Log Likelihood               -2075.827
Date:                Fri, 13 May 2022   AIC                           4159.654
Time:                        12:02:53   BIC                           4173.354
Sample:                    04-01-1996   HQIC                          4165.182
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.8265      0.129      6.392      0.000       0.573       1.080
ma.L1         -0.8894      0.133     -6.668      0.000      -1.151      -0.628
ma.L2         -0.0040      0.026     -0.154      0.878      -0.056       0.047
sigma2      4.979e+06   1.93e+05     25.739      0.000     4.6e+06    5.36e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   8.58   Jarque-Bera (JB):             1536
3.84
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              26.13   Skew:                            -
3.26
Prob(H) (two-sided):                  0.00   Kurtosis:                         4
2.77
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 14918.447255302292
                               SARIMAX Results
==============================================================================
Dep. Variable:           San Jose, CA   No. Observations:                  229
Model:                 ARIMA(1, 2, 2)   Log Likelihood               -2028.436
Date:                Fri, 13 May 2022   AIC                           4064.872
Time:                        12:02:54   BIC                           4078.571
Sample:                    04-01-1996   HQIC                          4070.400
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2798      1.767      0.158      0.874      -3.183       3.743
ma.L1         -0.2897      1.781     -0.163      0.871      -3.780       3.201
ma.L2         -0.0132      0.017     -0.755      0.451      -0.047       0.021
sigma2      3.013e+06   1.32e+05     22.874      0.000    2.75e+06    3.27e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   3.40   Jarque-Bera (JB):               79
4.12
Prob(Q):                              0.07   Prob(JB):
0.00
Heteroskedasticity (H):              15.68   Skew:
```

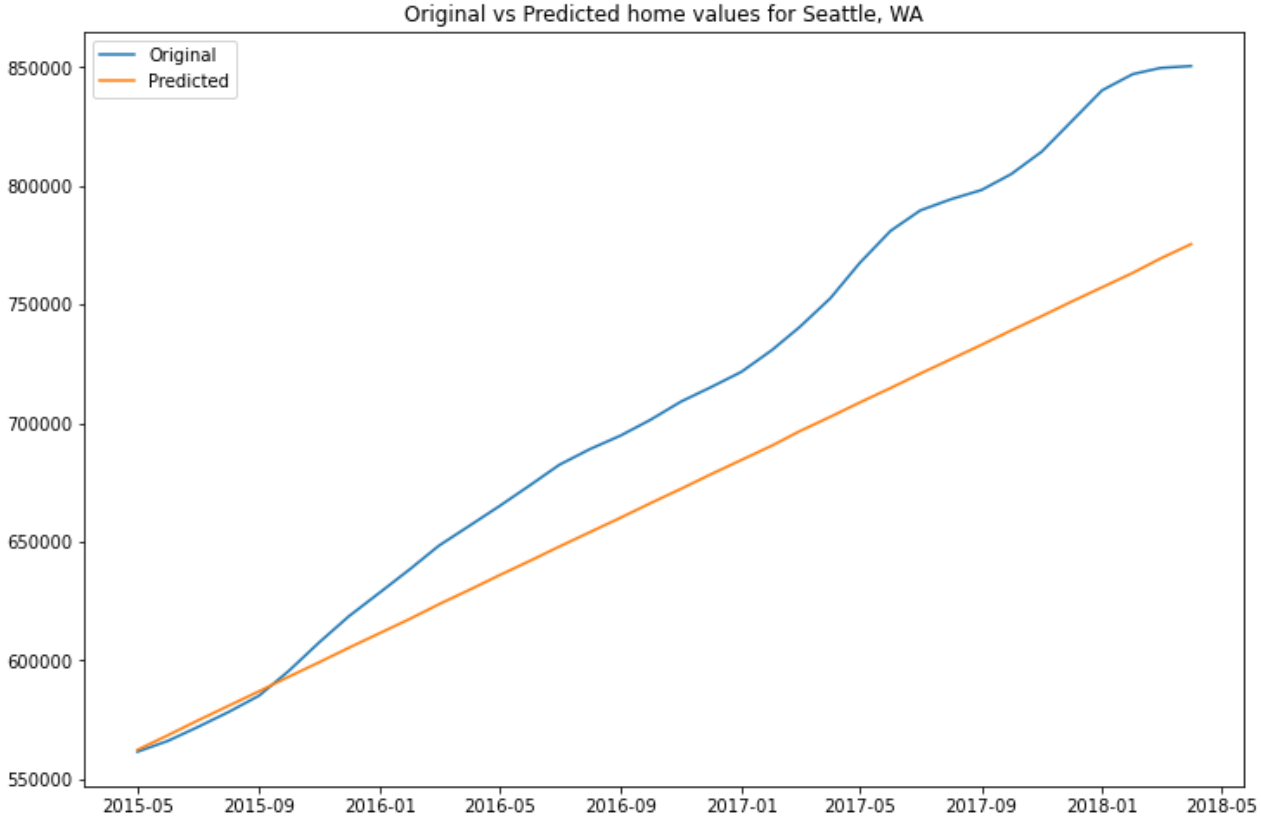
```
0.01
Prob(H) (two-sided):                  0.00   Kurtosis:                         1
2.16
===============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 33086.98593730578
                                SARIMAX Results
===============================================================================
Dep. Variable:            Chicago, IL   No. Observations:                  229
Model:                 ARIMA(1, 2, 2)   Log Likelihood               -1863.181
Date:                Fri, 13 May 2022   AIC                           3734.363
Time:                        12:02:54   BIC                           3748.063
Sample:                    04-01-1996   HQIC                          3739.891
                         - 04-01-2015
Covariance Type:                  opg
===============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
ar.L1          0.9324      0.194      4.807      0.000       0.552       1.313
ma.L1         -0.9221      0.192     -4.800      0.000      -1.299      -0.546
ma.L2         -0.0235      0.018     -1.279      0.201      -0.060       0.013
sigma2      7.831e+05   3.15e+04     24.853      0.000    7.21e+05    8.45e+05
===============================================================================
===
Ljung-Box (L1) (Q):                   9.78   Jarque-Bera (JB):                79
0.04
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              27.34   Skew:
0.55
Prob(H) (two-sided):                  0.00   Kurtosis:                         1
2.07
===============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 3903.5315642679493
                                SARIMAX Results
===============================================================================
Dep. Variable:          Baltimore, MD   No. Observations:                  229
Model:                 ARIMA(1, 2, 2)   Log Likelihood               -1746.510
Date:                Fri, 13 May 2022   AIC                           3501.020
Time:                        12:02:54   BIC                           3514.720
Sample:                    04-01-1996   HQIC                          3506.548
                         - 04-01-2015
Covariance Type:                  opg
===============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
ar.L1          0.3491      0.138      2.533      0.011       0.079       0.619
ma.L1         -0.3198      0.139     -2.299      0.022      -0.592      -0.047
ma.L2         -0.1896      0.021     -8.832      0.000      -0.232      -0.148
sigma2      2.452e+05   1.08e+04     22.806      0.000    2.24e+05    2.66e+05
===============================================================================
===
```

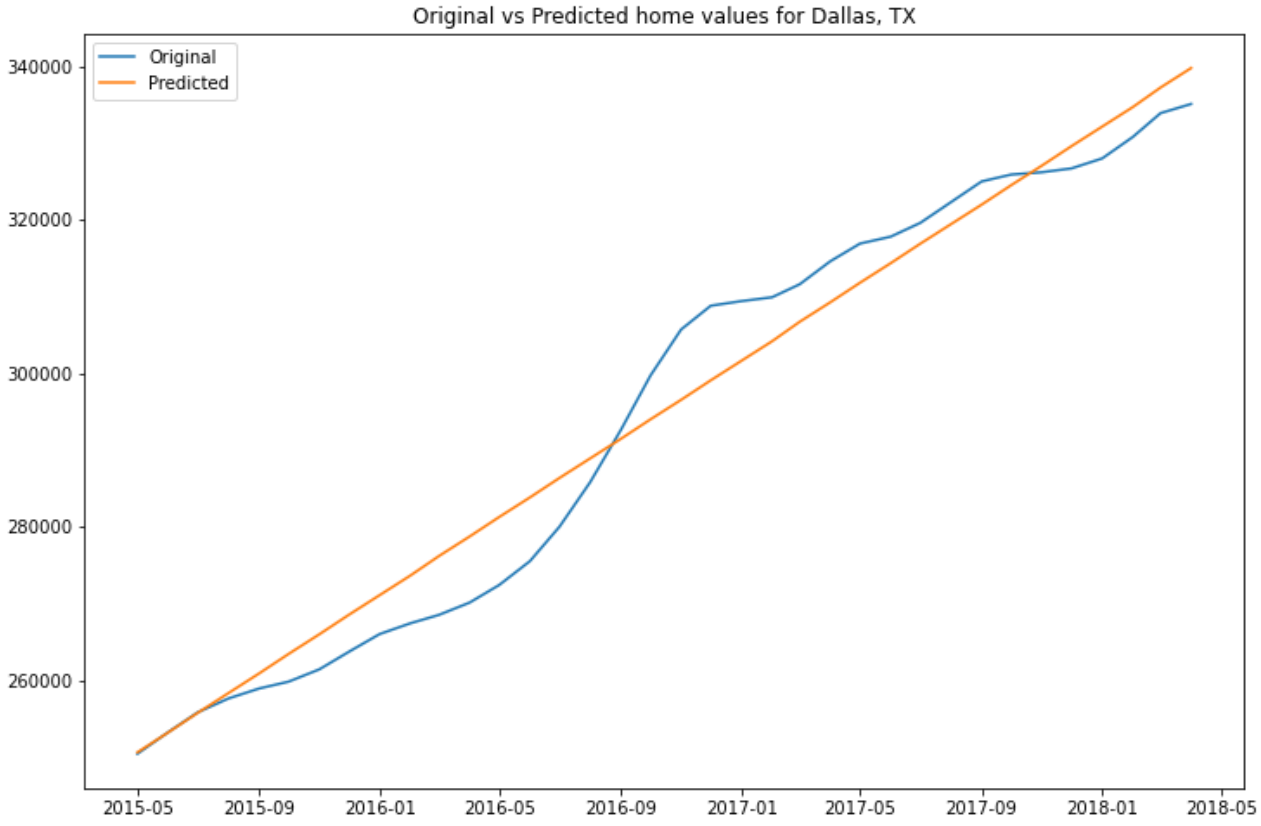
```
Ljung-Box (L1) (Q):                   2.09   Jarque-Bera (JB):              49
2.47
Prob(Q):                              0.15   Prob(JB):
0.00
Heteroskedasticity (H):              12.82   Skew:
1.13
Prob(H) (two-sided):                  0.00   Kurtosis:
9.86
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 16215.81153407634
                               SARIMAX Results
================================================================================
Dep. Variable:             Boston, MA   No. Observations:                  229
Model:                 ARIMA(1, 2, 2)   Log Likelihood               -1993.378
Date:                Fri, 13 May 2022   AIC                         3994.756
Time:                        12:02:55   BIC                         4008.456
Sample:                    04-01-1996   HQIC                        4000.285
                         - 04-01-2015
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.8093      0.098      8.245      0.000       0.617       1.002
ma.L1         -0.8958      0.092     -9.705      0.000      -1.077      -0.715
ma.L2         -0.0217      0.043     -0.504      0.614      -0.106       0.063
sigma2      2.482e+06   8.88e+04     27.943      0.000    2.31e+06    2.66e+06
================================================================================
===
Ljung-Box (L1) (Q):                   2.92   Jarque-Bera (JB):            1059
7.56
Prob(Q):                              0.09   Prob(JB):
0.00
Heteroskedasticity (H):              20.42   Skew:                          -
1.35
Prob(H) (two-sided):                  0.00   Kurtosis:                       3
6.36
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 52452.60559929399
```

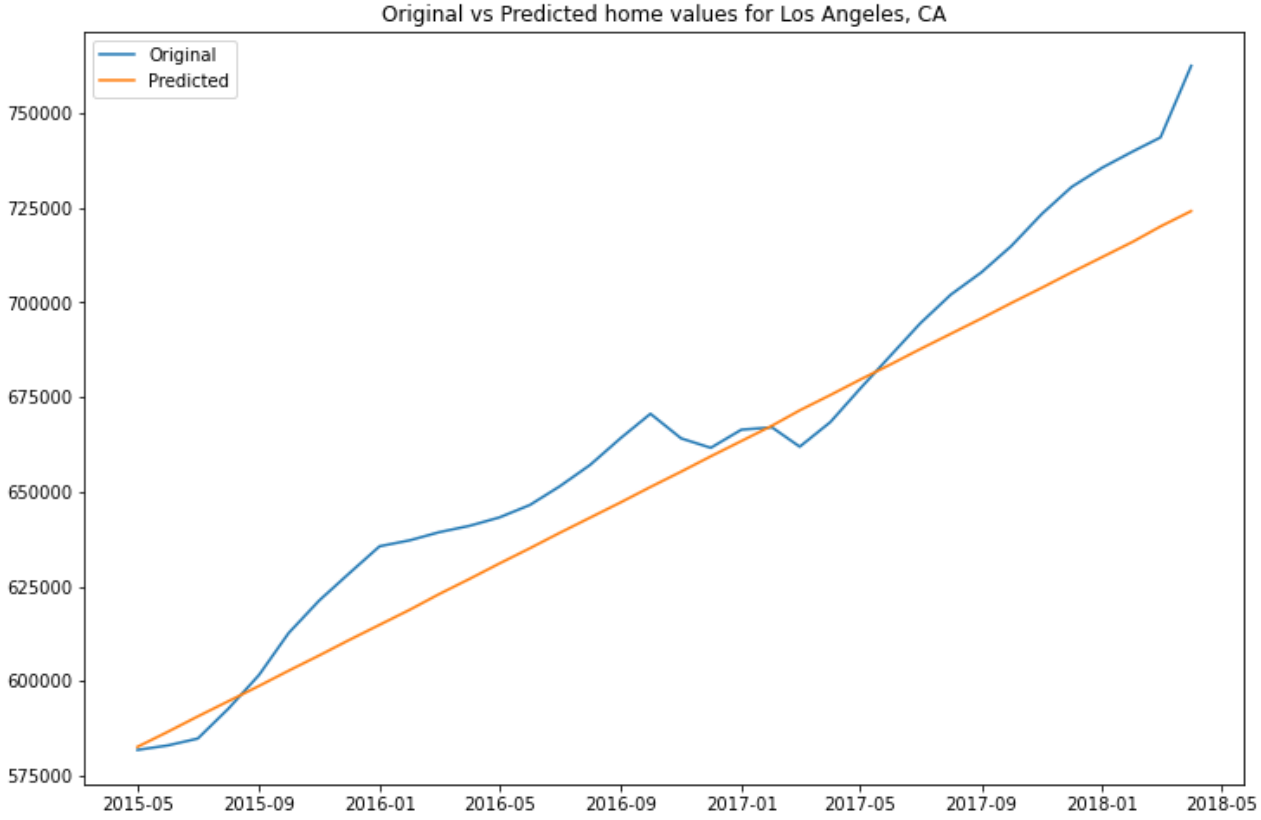

Original vs Predicted home values for Washington, DC
Original
Predicted
850000
800000
750000
700000
650000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

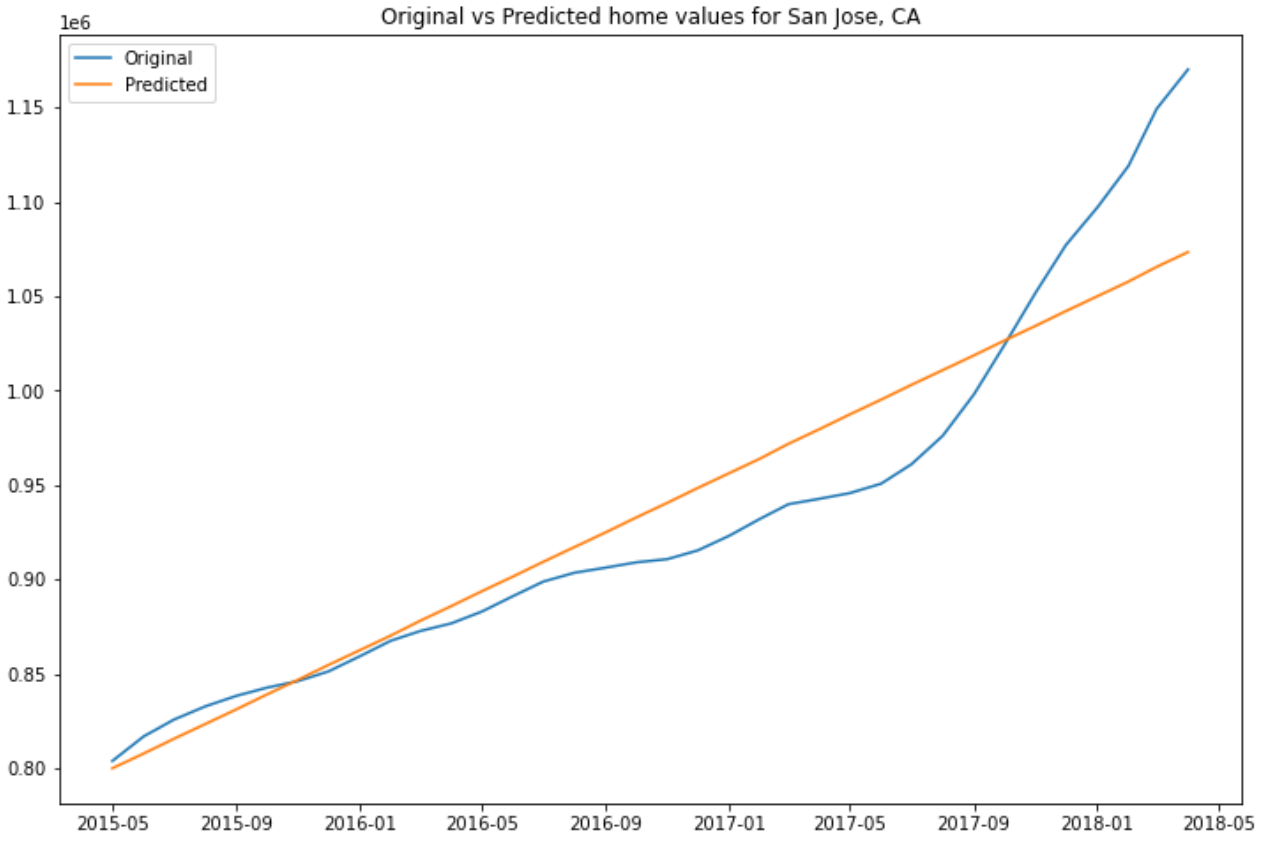

Original vs Predicted home values for New York, NY
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

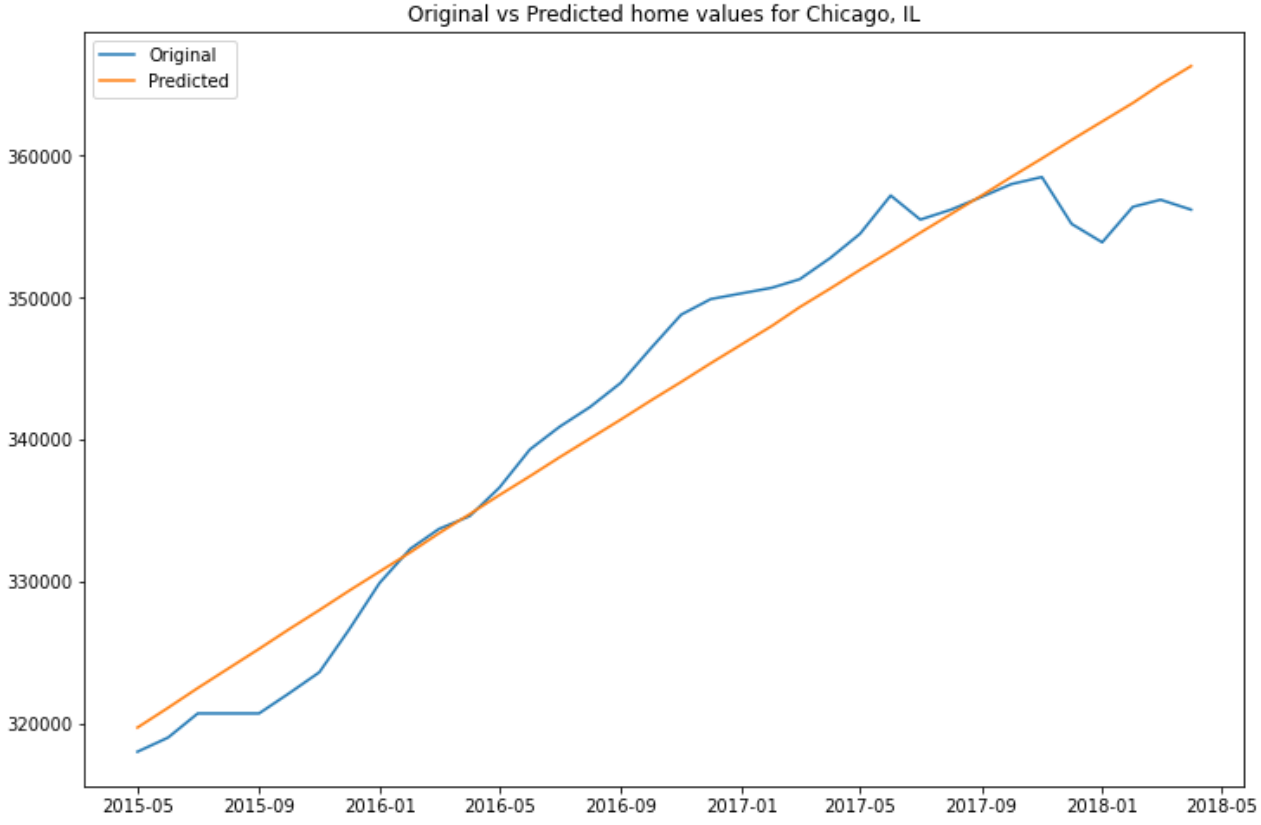

Original vs Predicted home values for San Francisco, CA
1e6
Original
Predicted
1.75
1.70
1.65
1.60
1.55
1.50
1.45
1.40
1.35
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

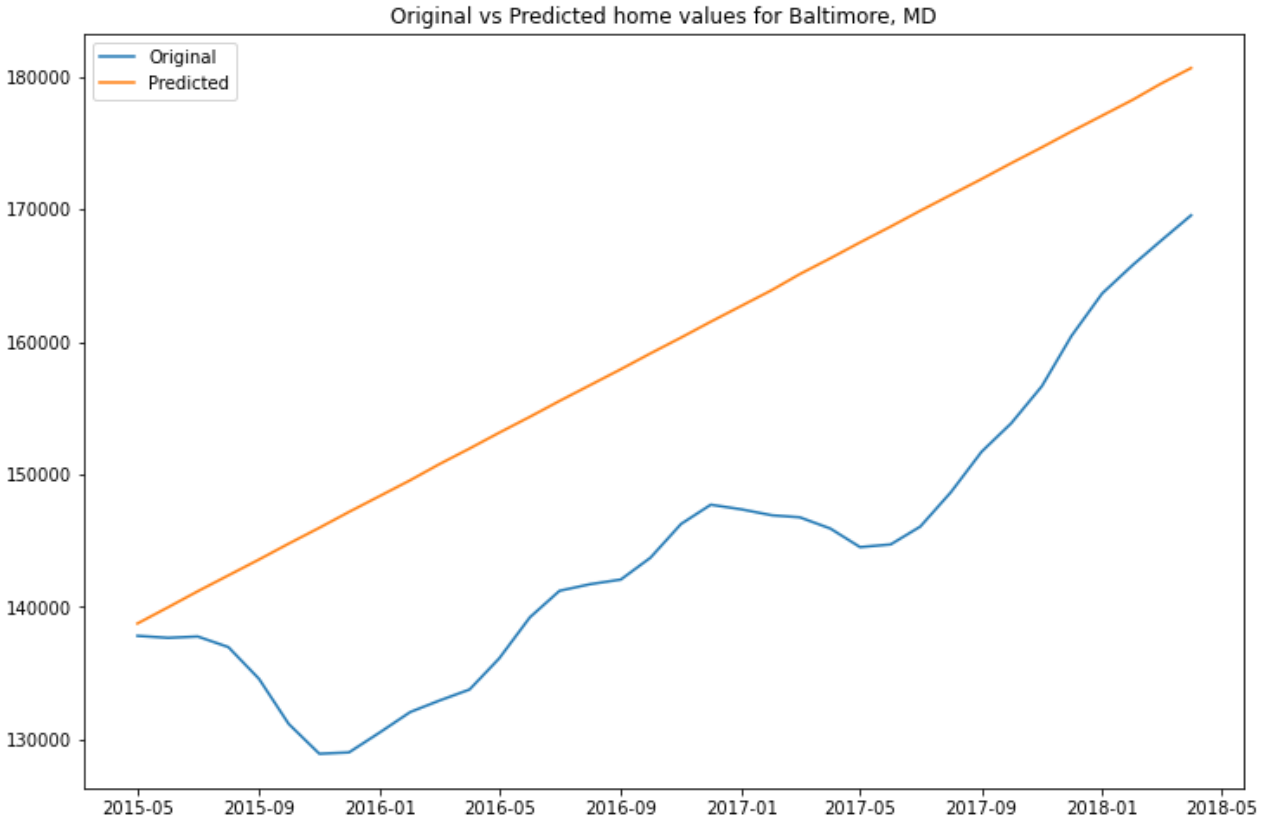

Original vs Predicted home values for Seattle, WA
Original
Predicted
850000
800000
750000
700000
650000
600000
550000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

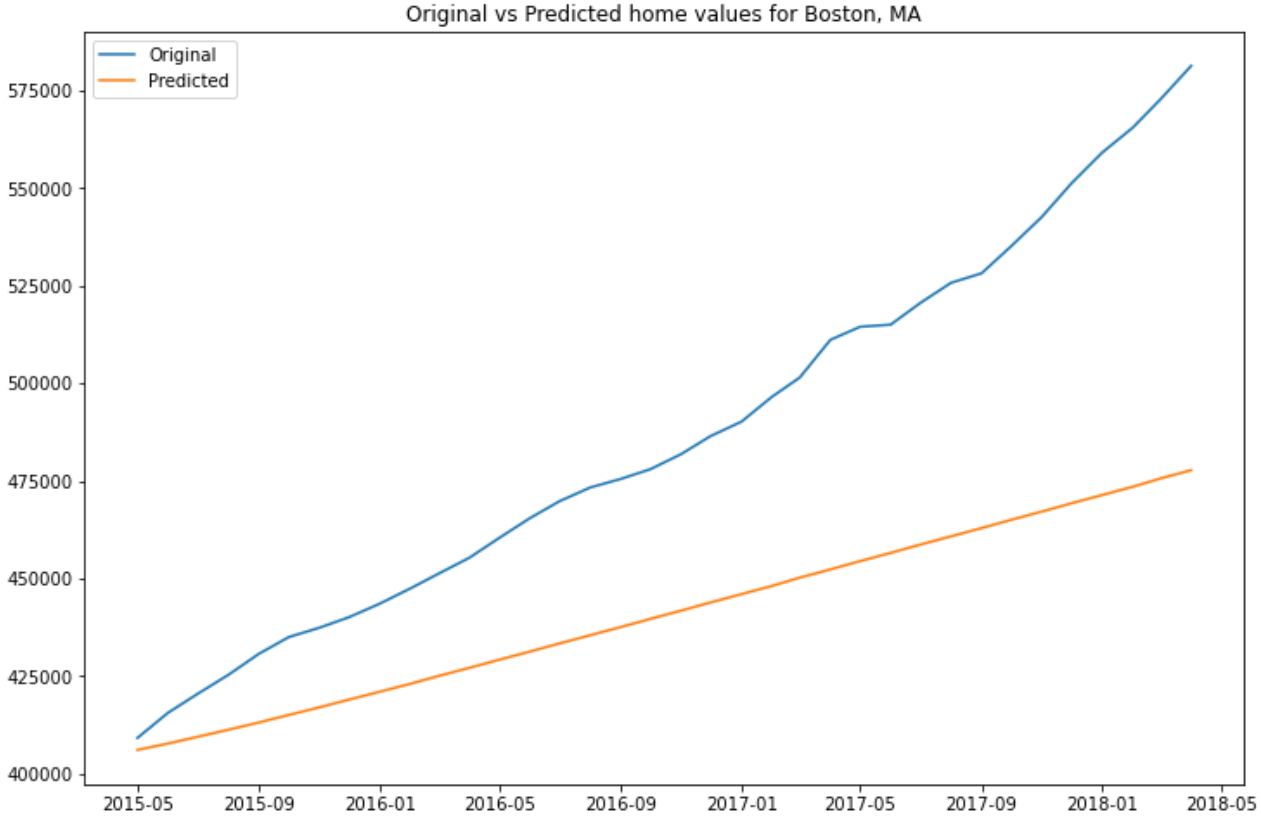

Original vs Predicted home values for Dallas, TX
Original
Predicted
340000
320000
300000
280000
260000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Los Angeles, CA
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for San Jose, CA
1e6
Original
Predicted
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Chicago, IL
Original
Predicted
360000
350000
340000
330000
320000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05





In [74]:
```
# even more improvement
np.mean(rmse_list)
```

Out[74]: 35695.0503309849

In [75]:
```
# see how adding in Autoregression helps
rmse_list = []
```

```
for city in city_list:
    city_model = arima_mod(city)
    city_model.model(train, test, 2,2,2)
    city_model.plot(test)
    rmse_list.append(city_model.rmse_)
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:          Washington, DC   No. Observations:                  229
Model:                 ARIMA(2, 2, 2)   Log Likelihood               -1858.448
Date:                Fri, 13 May 2022   AIC                           3726.897
Time:                        12:02:59   BIC                           3744.021
Sample:                    04-01-1996   HQIC                          3733.807
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3824      0.484      2.856      0.004       0.434       2.331
ar.L2         -0.4106      0.461     -0.891      0.373      -1.314       0.493
ma.L1         -1.3393      0.478     -2.803      0.005      -2.276      -0.403
ma.L2          0.3649      0.455      0.802      0.423      -0.527       1.257
sigma2      7.316e+05   4.56e+04     16.043      0.000    6.42e+05    8.21e+05
===================================================================================
Ljung-Box (L1) (Q):                  49.67   Jarque-Bera (JB):                 10
1.81
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              20.99   Skew:
0.09
Prob(H) (two-sided):                  0.00   Kurtosis:
6.28
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 45445.814636057716
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:978: UserWarning:

Non-invertible starting MA parameters found. Using zeros as starting parameters.

                               SARIMAX Results
==============================================================================
Dep. Variable:           New York, NY   No. Observations:                  229
Model:                 ARIMA(2, 2, 2)   Log Likelihood               -2127.546
Date:                Fri, 13 May 2022   AIC                           4265.093
Time:                        12:03:00   BIC                           4282.217
Sample:                    04-01-1996   HQIC                          4272.003
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3577      2.174      0.165      0.869      -3.904       4.619
ar.L2          0.4114      1.858      0.221      0.825      -3.231       4.054
ma.L1         -0.4960      2.163     -0.229      0.819      -4.736       3.744
ma.L2         -0.4875      2.138     -0.228      0.820      -4.678       3.703
sigma2      7.001e+06   4.96e-06   1.41e+12      0.000       7e+06       7e+06
```

```
==========================================================================
===
Ljung-Box (L1) (Q):                   1.87   Jarque-Bera (JB):               31
1.65
Prob(Q):                              0.17   Prob(JB):
0.00
Heteroskedasticity (H):              16.69   Skew:                            -
0.26
Prob(H) (two-sided):                  0.00   Kurtosis:
8.72
==========================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 1.11e+
28. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 62705.79342962915
                               SARIMAX Results
==========================================================================
Dep. Variable:      San Francisco, CA   No. Observations:                  229
Model:                 ARIMA(2, 2, 2)   Log Likelihood               -2144.626
Date:                Fri, 13 May 2022   AIC                         4299.252
Time:                        12:03:00   BIC                         4316.376
Sample:                    04-01-1996   HQIC                        4306.162
                         - 04-01-2015
Covariance Type:                  opg
==========================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------
ar.L1          1.2617      0.872      1.448      0.148      -0.447       2.970
ar.L2         -0.3060      0.822     -0.372      0.710      -1.918       1.306
ma.L1         -1.2903      0.862     -1.497      0.134      -2.979       0.399
ma.L2          0.2993      0.849      0.352      0.725      -1.366       1.964
sigma2      7.306e+06   2.47e-06   2.96e+12      0.000    7.31e+06    7.31e+06
==========================================================================
===
Ljung-Box (L1) (Q):                  21.16   Jarque-Bera (JB):                9
6.09
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              19.04   Skew:
0.43
Prob(H) (two-sided):                  0.00   Kurtosis:
6.07
==========================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 3.12e+
27. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 75466.71490089725
                               SARIMAX Results
==========================================================================
Dep. Variable:            Seattle, WA   No. Observations:                  229
Model:                 ARIMA(2, 2, 2)   Log Likelihood               -1883.180
Date:                Fri, 13 May 2022   AIC                         3776.361
Time:                        12:03:01   BIC                         3793.485
```

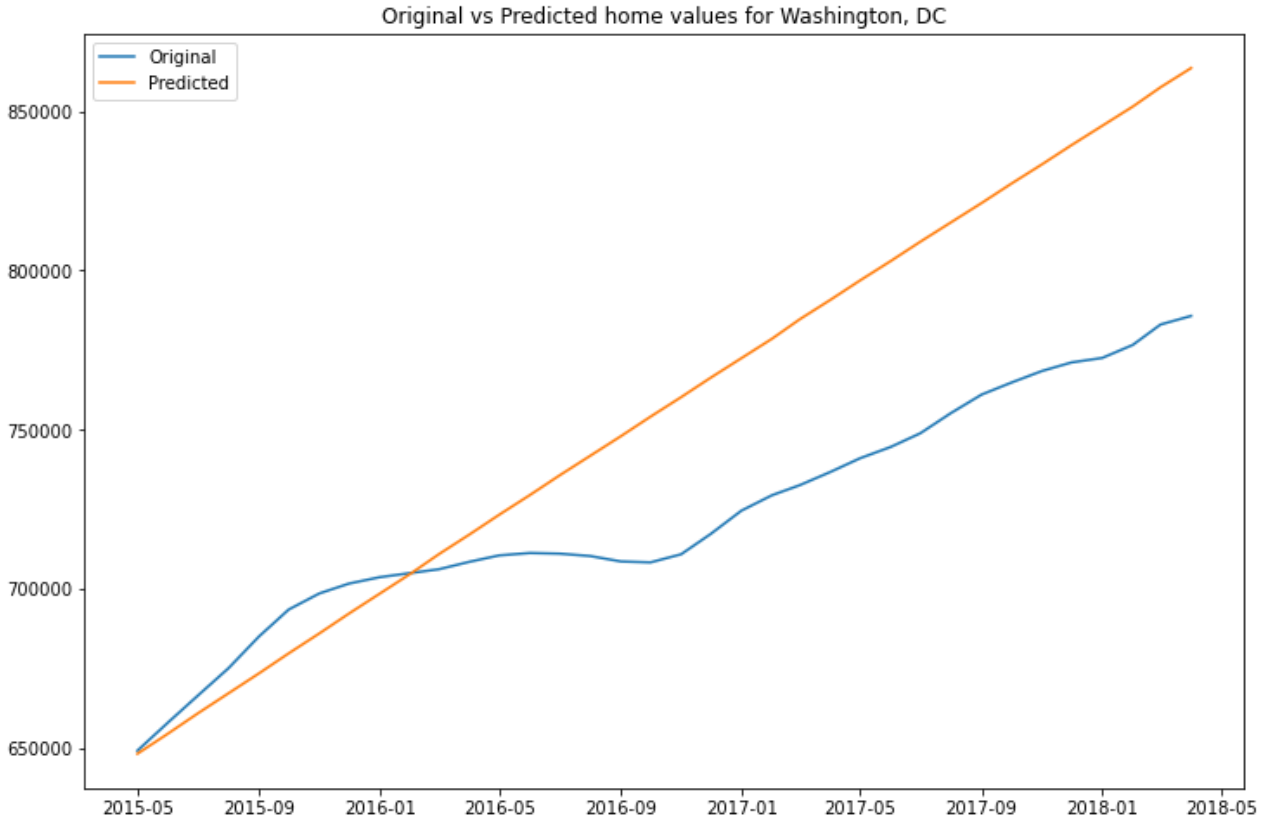
```
Sample:                     04-01-1996   HQIC                          3783.271
                          - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4047      0.114      3.560      0.000       0.182       0.628
ar.L2         -0.8553      0.099     -8.623      0.000      -1.050      -0.661
ma.L1         -0.3954      0.120     -3.306      0.001      -0.630      -0.161
ma.L2          0.8334      0.106      7.868      0.000       0.626       1.041
sigma2      6.194e+05   3.38e+04     18.320      0.000    5.53e+05    6.86e+05
==============================================================================
===
Ljung-Box (L1) (Q):                   24.56   Jarque-Bera (JB):                 1
9.98
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               17.50   Skew:
0.24
Prob(H) (two-sided):                  0.00   Kurtosis:
4.37
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 45240.86024832306
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:966: UserWarning:

Non-stationary starting autoregressive parameters found. Using zeros as starting
parameters.

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:978: UserWarning:

Non-invertible starting MA parameters found. Using zeros as starting parameters.

                               SARIMAX Results
==============================================================================
Dep. Variable:             Dallas, TX   No. Observations:                  229
Model:                 ARIMA(2, 2, 2)   Log Likelihood               -1951.922
Date:                Fri, 13 May 2022   AIC                           3913.844
Time:                        12:03:02   BIC                           3930.969
Sample:                     04-01-1996   HQIC                          3920.754
                          - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.7327      1.698      1.020      0.308      -1.596       5.061
ar.L2         -0.7333      1.353     -0.542      0.588      -3.384       1.918
ma.L1         -1.7653      1.697     -1.040      0.298      -5.092       1.561
ma.L2          0.7658      1.402      0.546      0.585      -1.982       3.514
sigma2      1.719e+06   3.46e-05   4.97e+10      0.000    1.72e+06    1.72e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   38.79   Jarque-Bera (JB):             7677
2.00
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               43.43   Skew:                             -
```

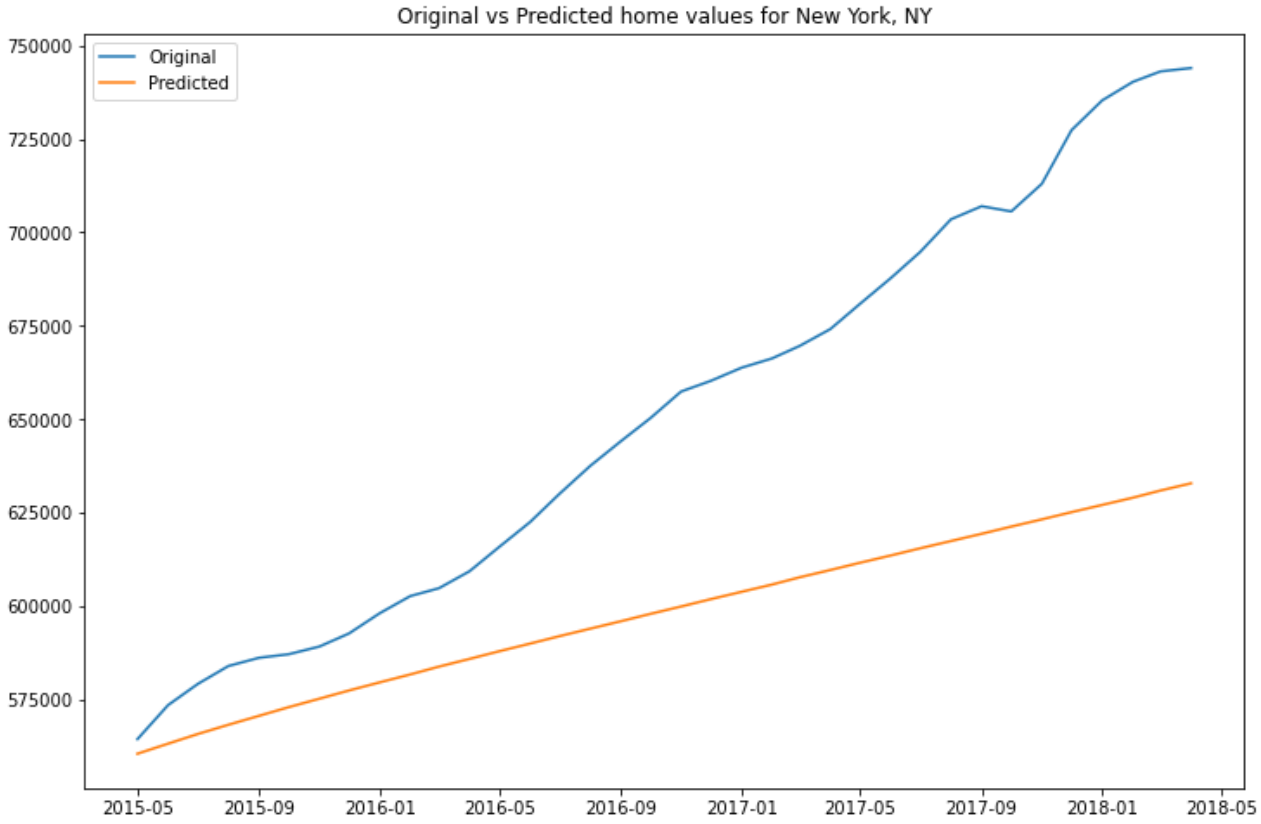
```
1.50
Prob(H) (two-sided):                  0.00   Kurtosis:                         9
3.04
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 1.67e+
26. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 5181.878353888262
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:966: UserWarning:

Non-stationary starting autoregressive parameters found. Using zeros as starting
parameters.

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:978: UserWarning:

Non-invertible starting MA parameters found. Using zeros as starting parameters.

                               SARIMAX Results
================================================================================
Dep. Variable:        Los Angeles, CA   No. Observations:                  229
Model:                 ARIMA(2, 2, 2)   Log Likelihood               -2075.851
Date:                Fri, 13 May 2022   AIC                           4161.701
Time:                        12:03:02   BIC                           4178.826
Sample:                    04-01-1996   HQIC                          4168.612
                         - 04-01-2015
Covariance Type:                  opg
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
ar.L1          0.0770      8.198      0.009      0.993     -15.991      16.145
ar.L2          0.6546      6.883      0.095      0.924     -12.836      14.146
ma.L1         -0.1399      8.197     -0.017      0.986     -16.206      15.926
ma.L2         -0.7018      7.408     -0.095      0.925     -15.221      13.818
sigma2      4.965e+06   1.93e+05     25.763      0.000    4.59e+06    5.34e+06
================================================================================
===
Ljung-Box (L1) (Q):                   8.53   Jarque-Bera (JB):             1535
2.09
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              26.74   Skew:                            -
3.25
Prob(H) (two-sided):                  0.00   Kurtosis:                         4
2.76
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
-----------------------
-----------------------
RMSE: 15037.296175161033
                               SARIMAX Results
================================================================================
Dep. Variable:           San Jose, CA   No. Observations:                  229
Model:                 ARIMA(2, 2, 2)   Log Likelihood               -2028.846
```

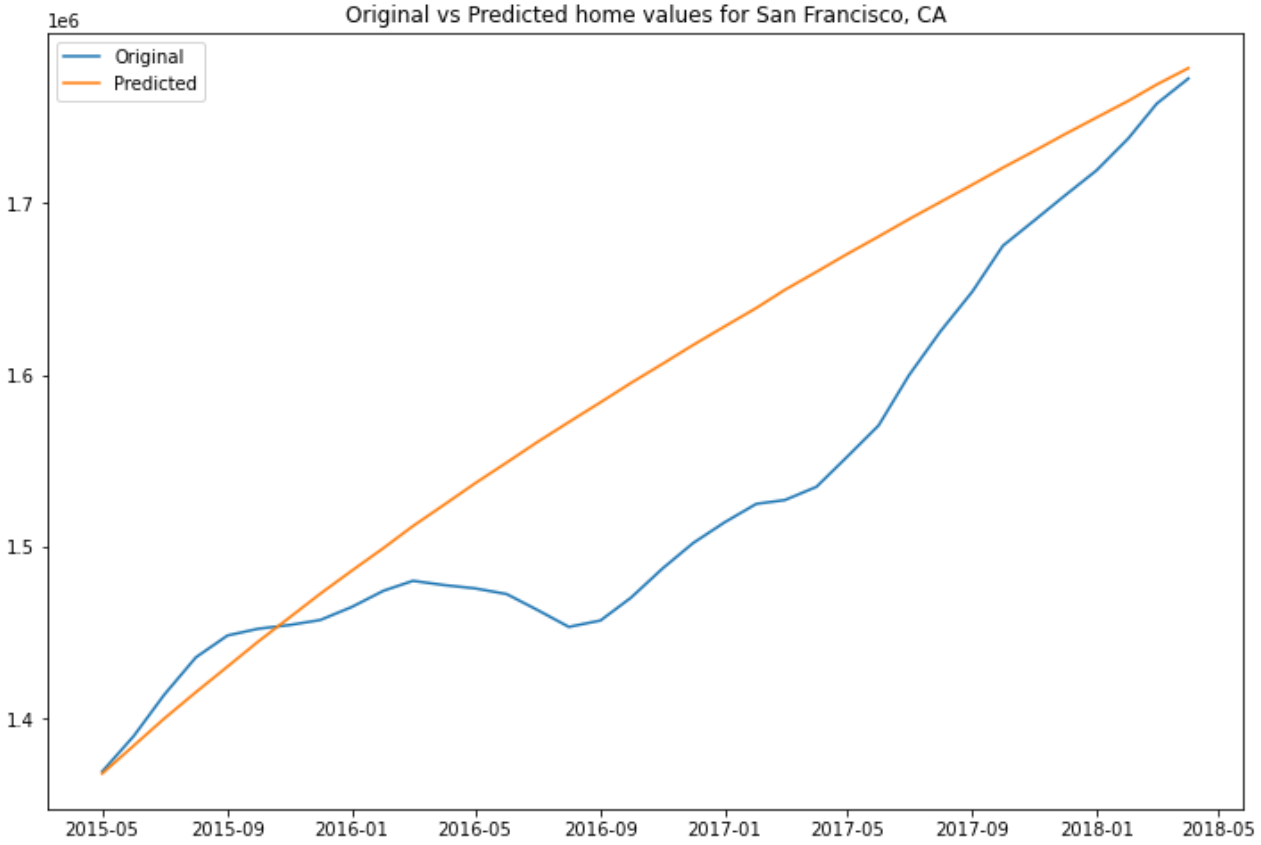
```
Date:                Fri, 13 May 2022   AIC                           4067.692
Time:                        12:03:03   BIC                           4084.816
Sample:                  04-01-1996   HQIC                          4074.602
                       - 04-01-2015
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2549      2.115      0.121      0.904      -3.891       4.401
ar.L2         -0.1525      2.411     -0.063      0.950      -4.877       4.572
ma.L1         -0.2640      2.128     -0.124      0.901      -4.435       3.907
ma.L2          0.1395      2.426      0.057      0.954      -4.616       4.895
sigma2       2.92e+06   1.26e+05     23.147      0.000    2.67e+06    3.17e+06
===================================================================================
===
Ljung-Box (L1) (Q):                   3.36   Jarque-Bera (JB):                79
4.85
Prob(Q):                              0.07   Prob(JB):
0.00
Heteroskedasticity (H):              15.62   Skew:
0.00
Prob(H) (two-sided):                  0.00   Kurtosis:                         1
2.17
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
-----------------------
-----------------------
RMSE: 33090.53470455137
                               SARIMAX Results
==============================================================================
Dep. Variable:            Chicago, IL   No. Observations:                  229
Model:                 ARIMA(2, 2, 2)   Log Likelihood               -1863.169
Date:                Fri, 13 May 2022   AIC                           3736.339
Time:                        12:03:03   BIC                           3753.463
Sample:                  04-01-1996   HQIC                          3743.249
                       - 04-01-2015
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9994      1.040      0.961      0.336      -1.039       3.037
ar.L2         -0.0688      1.016     -0.068      0.946      -2.060       1.923
ma.L1         -0.9883      1.036     -0.954      0.340      -3.018       1.042
ma.L2          0.0449      1.018      0.044      0.965      -1.951       2.041
sigma2      7.837e+05   3.19e+04     24.541      0.000    7.21e+05    8.46e+05
===================================================================================
===
Ljung-Box (L1) (Q):                   9.74   Jarque-Bera (JB):                79
0.89
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              27.34   Skew:
0.55
Prob(H) (two-sided):                  0.00   Kurtosis:                         1
2.08
===================================================================================
===

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

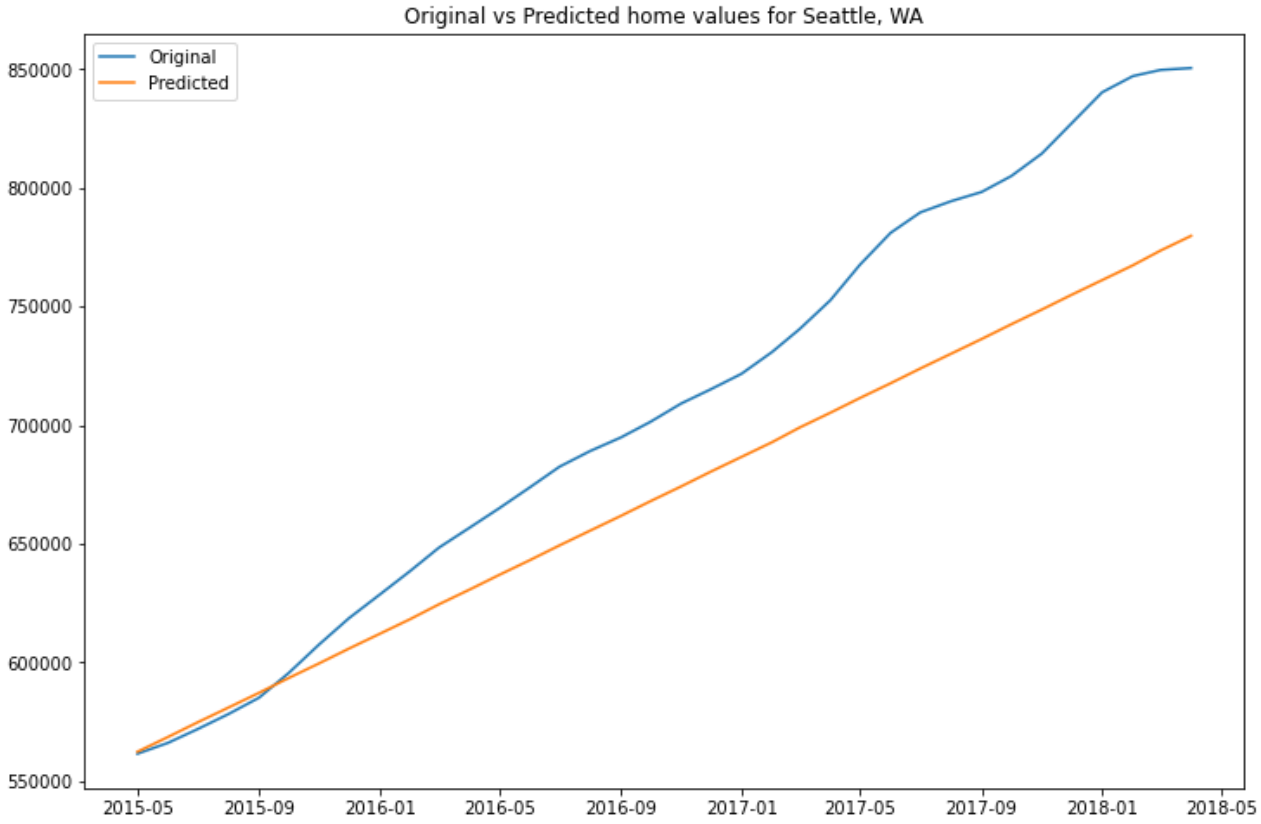
```
-----------------------
-----------------------
RMSE: 3970.9376439330927
                               SARIMAX Results
==============================================================================
Dep. Variable:          Baltimore, MD   No. Observations:                  229
Model:                 ARIMA(2, 2, 2)   Log Likelihood               -1746.485
Date:                Fri, 13 May 2022   AIC                           3502.970
Time:                        12:03:04   BIC                           3520.095
Sample:                    04-01-1996   HQIC                          3509.880
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3301      0.147      2.238      0.025       0.041       0.619
ar.L2          0.0492      0.168      0.293      0.769      -0.280       0.378
ma.L1         -0.3006      0.149     -2.018      0.044      -0.593      -0.009
ma.L2         -0.2391      0.172     -1.394      0.163      -0.575       0.097
sigma2      2.452e+05    1.1e+04     22.204      0.000    2.24e+05    2.67e+05
==============================================================================
===
Ljung-Box (L1) (Q):                   2.08   Jarque-Bera (JB):                 48
8.54
Prob(Q):                              0.15   Prob(JB):
0.00
Heteroskedasticity (H):              12.75   Skew:
1.13
Prob(H) (two-sided):                  0.00   Kurtosis:
9.83
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
-----------------------
-----------------------
RMSE: 16217.652497914098
                               SARIMAX Results
==============================================================================
Dep. Variable:              Boston, MA   No. Observations:                  229
Model:                 ARIMA(2, 2, 2)   Log Likelihood               -1993.771
Date:                Fri, 13 May 2022   AIC                           3997.542
Time:                        12:03:04   BIC                           4014.667
Sample:                    04-01-1996   HQIC                          4004.452
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1146      0.596     -0.192      0.848      -1.282       1.053
ar.L2          0.8288      0.498      1.665      0.096      -0.147       1.804
ma.L1          0.0250      0.594      0.042      0.966      -1.138       1.188
ma.L2         -0.9140      0.554     -1.651      0.099      -1.999       0.171
sigma2      2.408e+06    8.2e+04     29.360      0.000    2.25e+06    2.57e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   2.55   Jarque-Bera (JB):              1028
9.93
Prob(Q):                              0.11   Prob(JB):
0.00
Heteroskedasticity (H):              22.15   Skew:                             -
1.27
Prob(H) (two-sided):                  0.00   Kurtosis:                          3
```

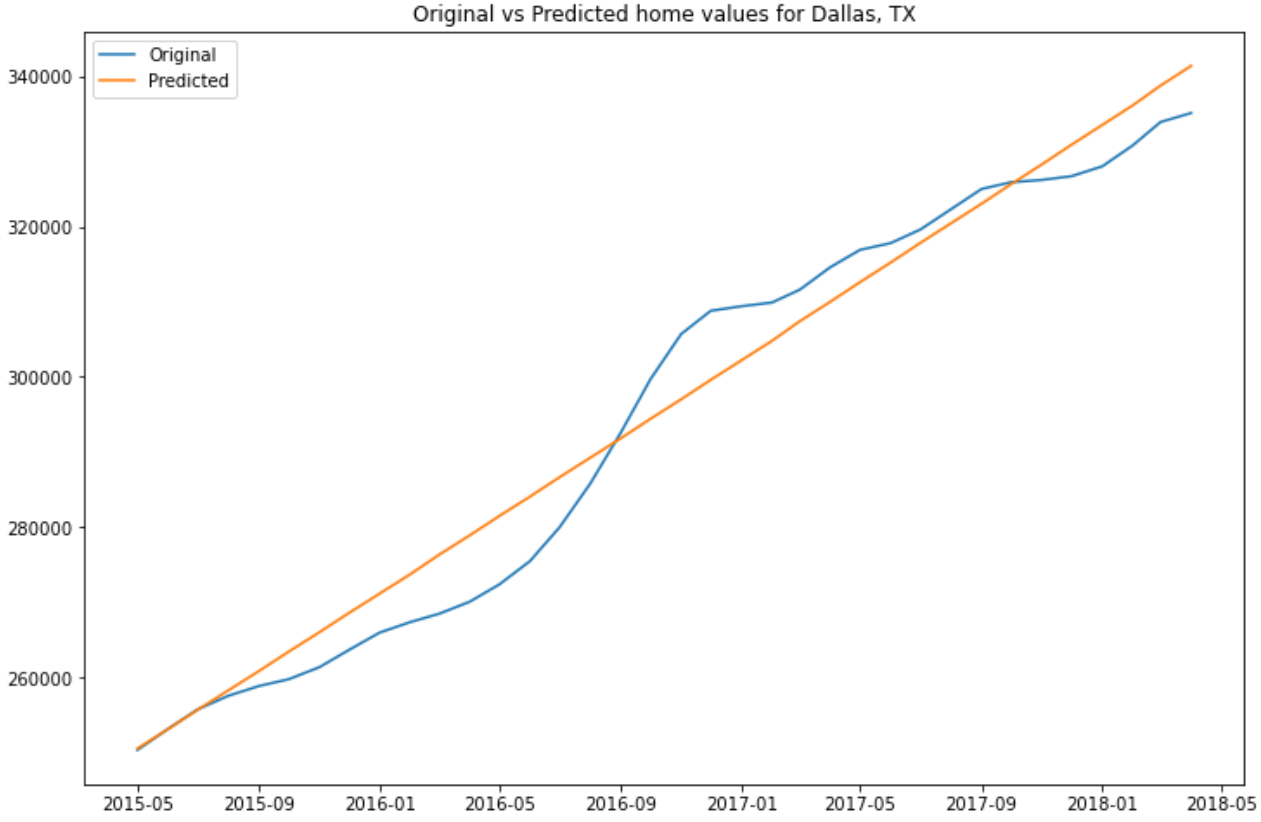
```
5.89
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 54310.08442006778
```

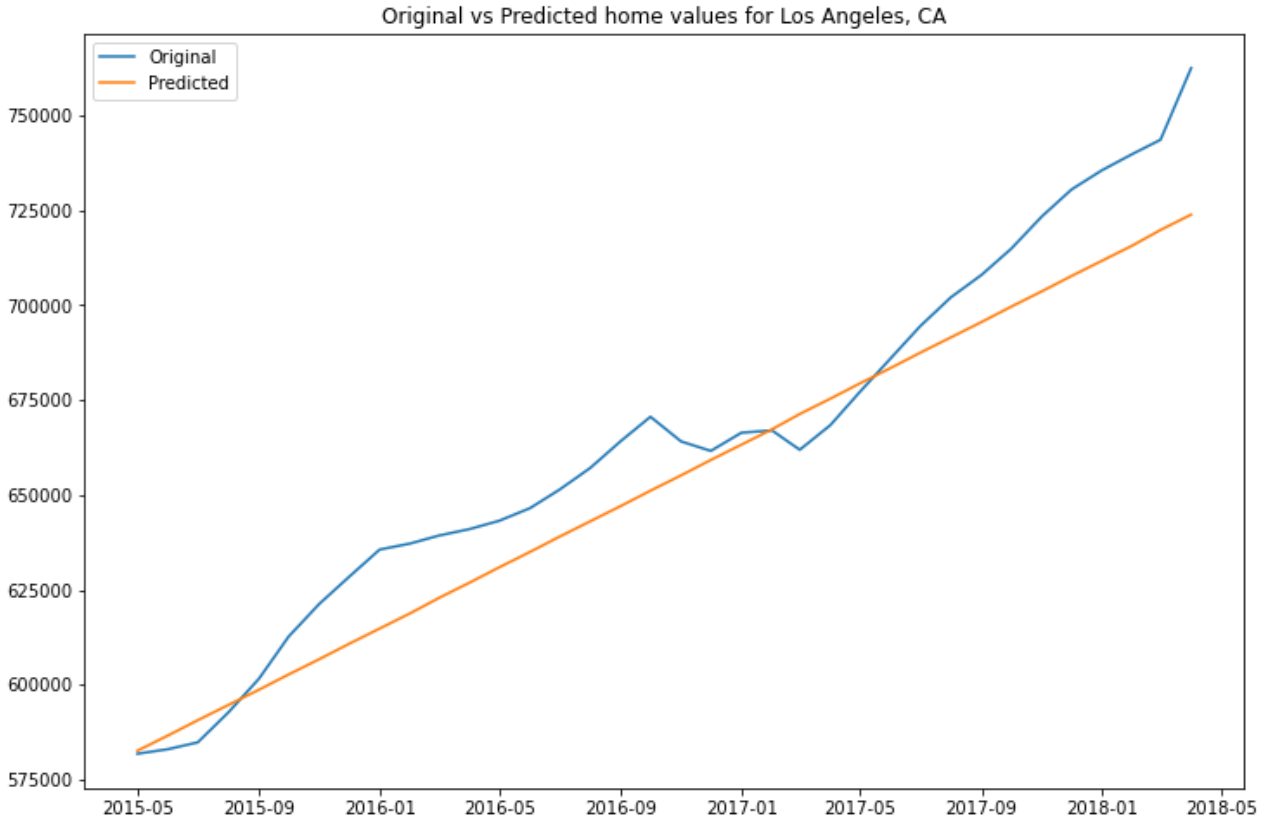


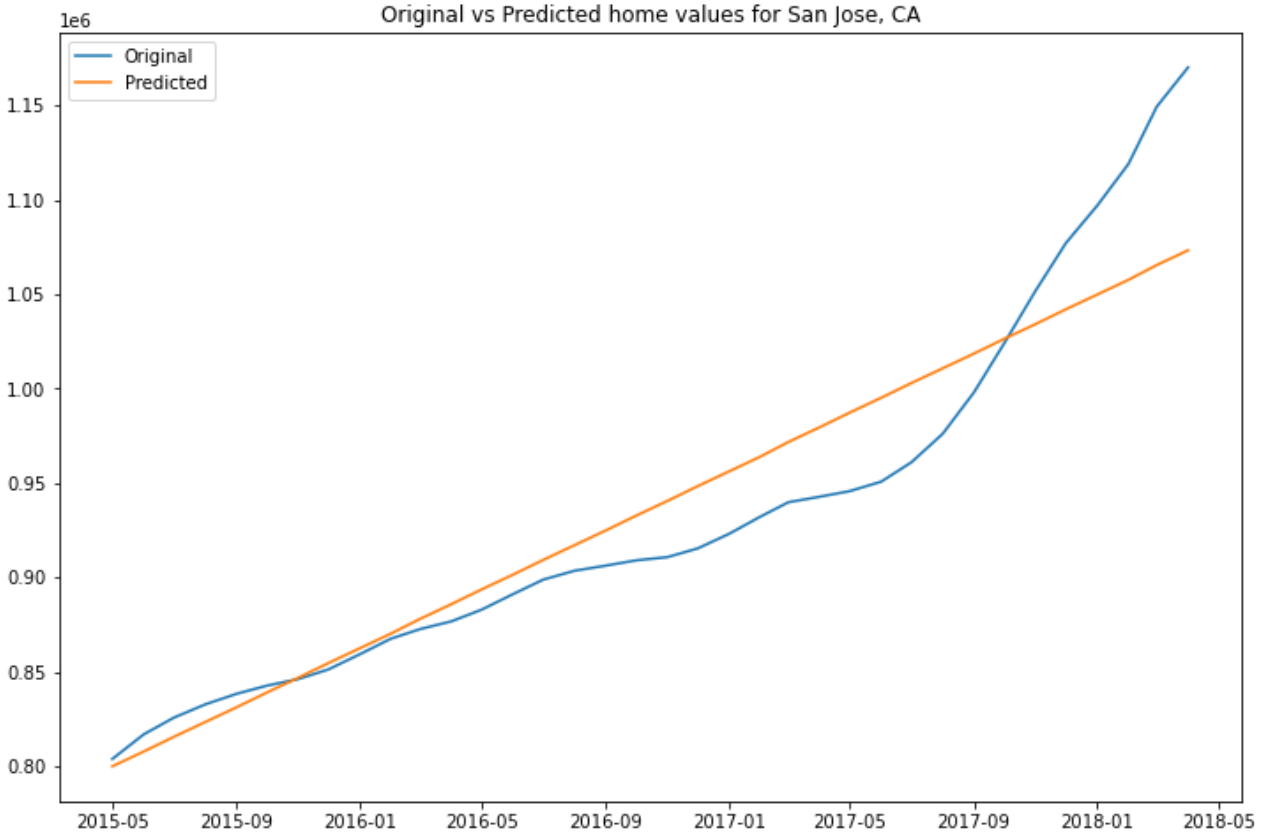

Original vs Predicted home values for New York, NY
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

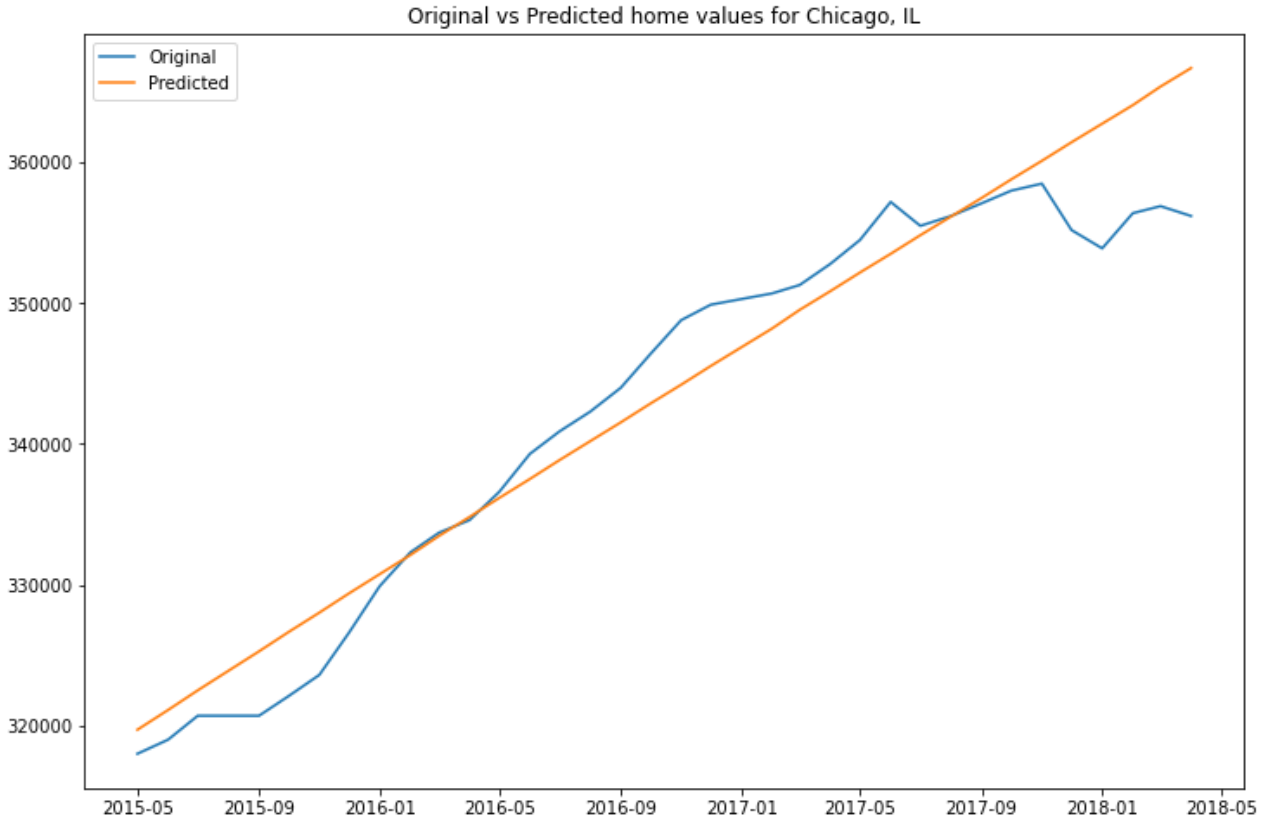

Original vs Predicted home values for San Francisco, CA
1e6
Original
Predicted
1.7
1.6
1.5
1.4
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

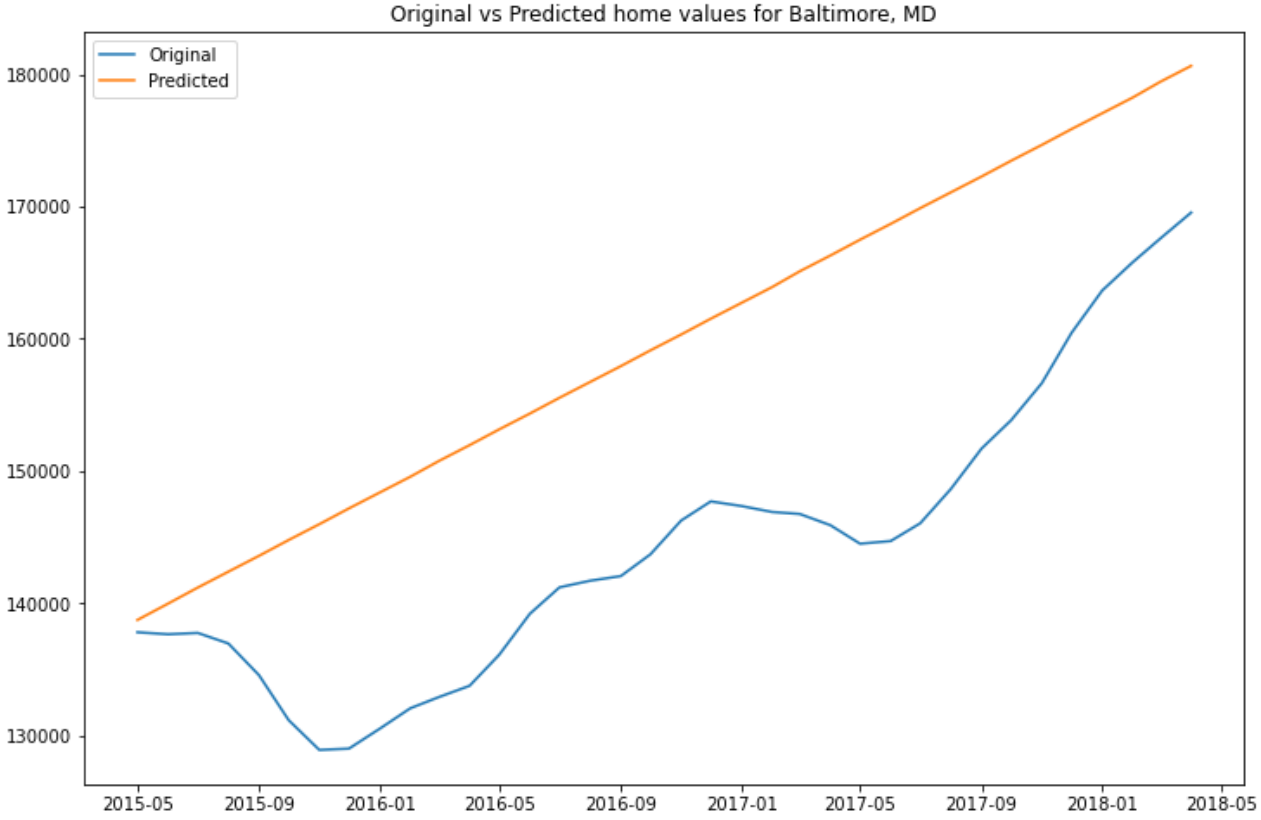

Original vs Predicted home values for Seattle, WA
Original
Predicted
850000
800000
750000
700000
650000
600000
550000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

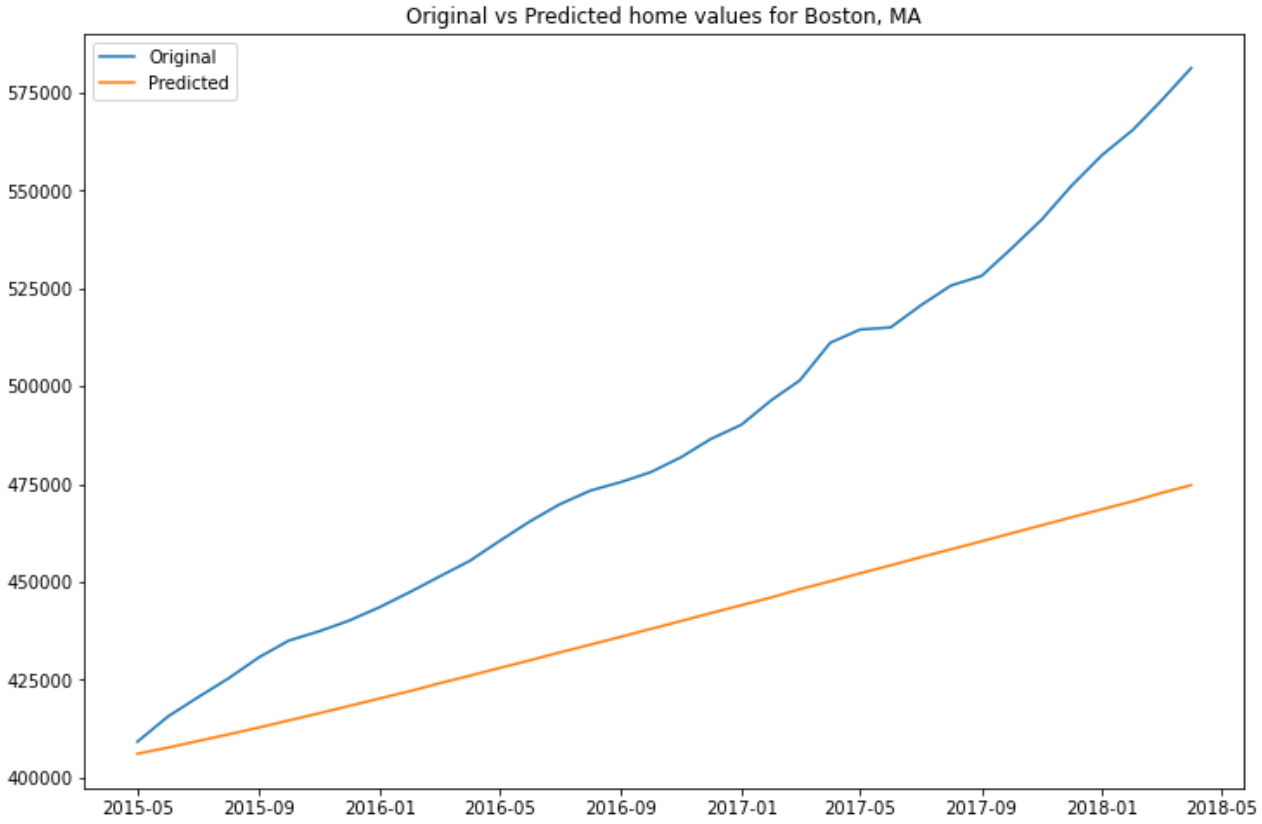

Original vs Predicted home values for Dallas, TX
Original
Predicted
340000
320000
300000
280000
260000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Los Angeles, CA
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for San Jose, CA
1e6
Original
Predicted
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Chicago, IL
Original
Predicted
360000
350000
340000
330000
320000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Baltimore, MD
Original
Predicted
180000
170000
160000
150000
140000
130000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05



In [76]:
```
# moving in the wrong direction
np.mean(rmse_list)
```

Out[76]: 35666.75670104228

In [77]:
```
# perhaps giving another moving average variable
rmse_list = []
for city in city_list:
    city_model = arima_mod(city)
    city_model.model(train, test, 1,2,3)
    city_model.plot(test)
    rmse_list.append(city_model.rmse_)
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:          Washington, DC   No. Observations:                  229
Model:                  ARIMA(1, 2, 3)   Log Likelihood               -1857.847
Date:                 Fri, 13 May 2022   AIC                           3725.693
Time:                         12:03:09   BIC                           3742.818
Sample:                     04-01-1996   HQIC                          3732.603
                          - 04-01-2015
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9714      0.023     41.616      0.000       0.926       1.017
ma.L1         -0.9290      0.022    -41.750      0.000      -0.973      -0.885
ma.L2         -0.0345      0.017     -2.056      0.040      -0.067      -0.002
ma.L3         -0.0238      0.024     -1.006      0.314      -0.070       0.023
sigma2      7.277e+05   4.39e+04     16.594      0.000    6.42e+05    8.14e+05
==============================================================================
===
Ljung-Box (L1) (Q):                   49.02   Jarque-Bera (JB):                 10
3.80
Prob(Q):                              0.00   Prob(JB):
```

```
0.00
Heteroskedasticity (H):              15.74   Skew:
0.25
Prob(H) (two-sided):                  0.00   Kurtosis:
6.28
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 36748.34494860473
                               SARIMAX Results
================================================================================
Dep. Variable:           New York, NY   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -2126.309
Date:                Fri, 13 May 2022   AIC                          4262.617
Time:                        12:03:10   BIC                          4279.742
Sample:                    04-01-1996   HQIC                         4269.528
                         - 04-01-2015
Covariance Type:                  opg
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
ar.L1          0.7494      0.076      9.887      0.000       0.601       0.898
ma.L1         -0.8831      0.093     -9.498      0.000      -1.065      -0.701
ma.L2         -0.0310      0.059     -0.530      0.596      -0.146       0.084
ma.L3         -0.0220      0.027     -0.806      0.420      -0.075       0.031
sigma2      7.341e+06    3.9e+05     18.834      0.000    6.58e+06     8.1e+06
================================================================================
===
Ljung-Box (L1) (Q):                   1.55   Jarque-Bera (JB):               27
8.23
Prob(Q):                              0.21   Prob(JB):
0.00
Heteroskedasticity (H):              14.04   Skew:                            -
0.21
Prob(H) (two-sided):                  0.00   Kurtosis:
8.41
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 50867.36839003728
                               SARIMAX Results
================================================================================
Dep. Variable:      San Francisco, CA   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -2142.881
Date:                Fri, 13 May 2022   AIC                          4295.762
Time:                        12:03:10   BIC                          4312.887
Sample:                    04-01-1996   HQIC                         4302.672
                         - 04-01-2015
Covariance Type:                  opg
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
ar.L1          0.9388      0.025     38.299      0.000       0.891       0.987
ma.L1         -0.9759      0.045    -21.771      0.000      -1.064      -0.888
ma.L2         -0.0160      0.011     -1.483      0.138      -0.037       0.005
```

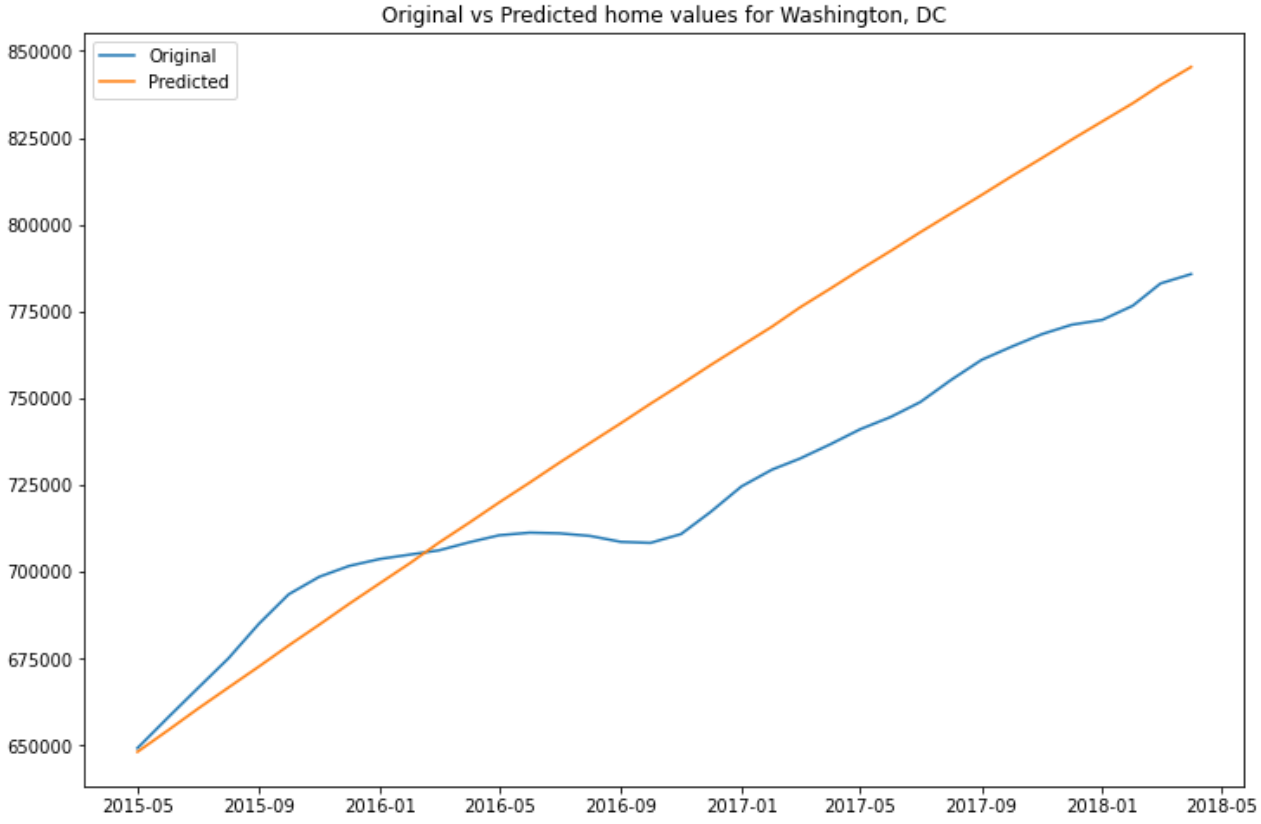
```
ma.L3         -0.0070      0.025     -0.280      0.780      -0.056       0.042
sigma2      7.765e+06   2.85e-09   2.73e+15      0.000    7.77e+06    7.77e+06
===================================================================================
===
Ljung-Box (L1) (Q):                   21.12   Jarque-Bera (JB):                 9
1.84
Prob(Q):                               0.00   Prob(JB):
0.00
Heteroskedasticity (H):               18.69   Skew:
0.40
Prob(H) (two-sided):                   0.00   Kurtosis:
6.01
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 1.27e+
30. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 67963.99107969219
                               SARIMAX Results
===================================================================================
Dep. Variable:            Seattle, WA   No. Observations:                   229
Model:                 ARIMA(1, 2, 3)   Log Likelihood                -1878.103
Date:                Fri, 13 May 2022   AIC                            3766.205
Time:                        12:03:11   BIC                            3783.330
Sample:                    04-01-1996   HQIC                           3773.115
                         - 04-01-2015
Covariance Type:                  opg
===================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
ar.L1         -0.0221      2.478     -0.009      0.993      -4.879       4.835
ma.L1          0.0281      2.484      0.011      0.991      -4.840       4.896
ma.L2         -0.0146      0.036     -0.411      0.681      -0.084       0.055
ma.L3         -0.0168      0.025     -0.665      0.506      -0.066       0.033
sigma2      8.889e+05   6.53e+04     13.610      0.000    7.61e+05    1.02e+06
===================================================================================
===
Ljung-Box (L1) (Q):                   24.43   Jarque-Bera (JB):                 2
5.50
Prob(Q):                               0.00   Prob(JB):
0.00
Heteroskedasticity (H):               26.23   Skew:
0.28
Prob(H) (two-sided):                   0.00   Kurtosis:
4.54
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
-----------------------
-----------------------
RMSE: 45483.41561208202
                               SARIMAX Results
===================================================================================
Dep. Variable:              Dallas, TX   No. Observations:                   229
Model:                 ARIMA(1, 2, 3)   Log Likelihood                -1955.950
Date:                Fri, 13 May 2022   AIC                            3921.900
Time:                        12:03:11   BIC                            3939.025
```

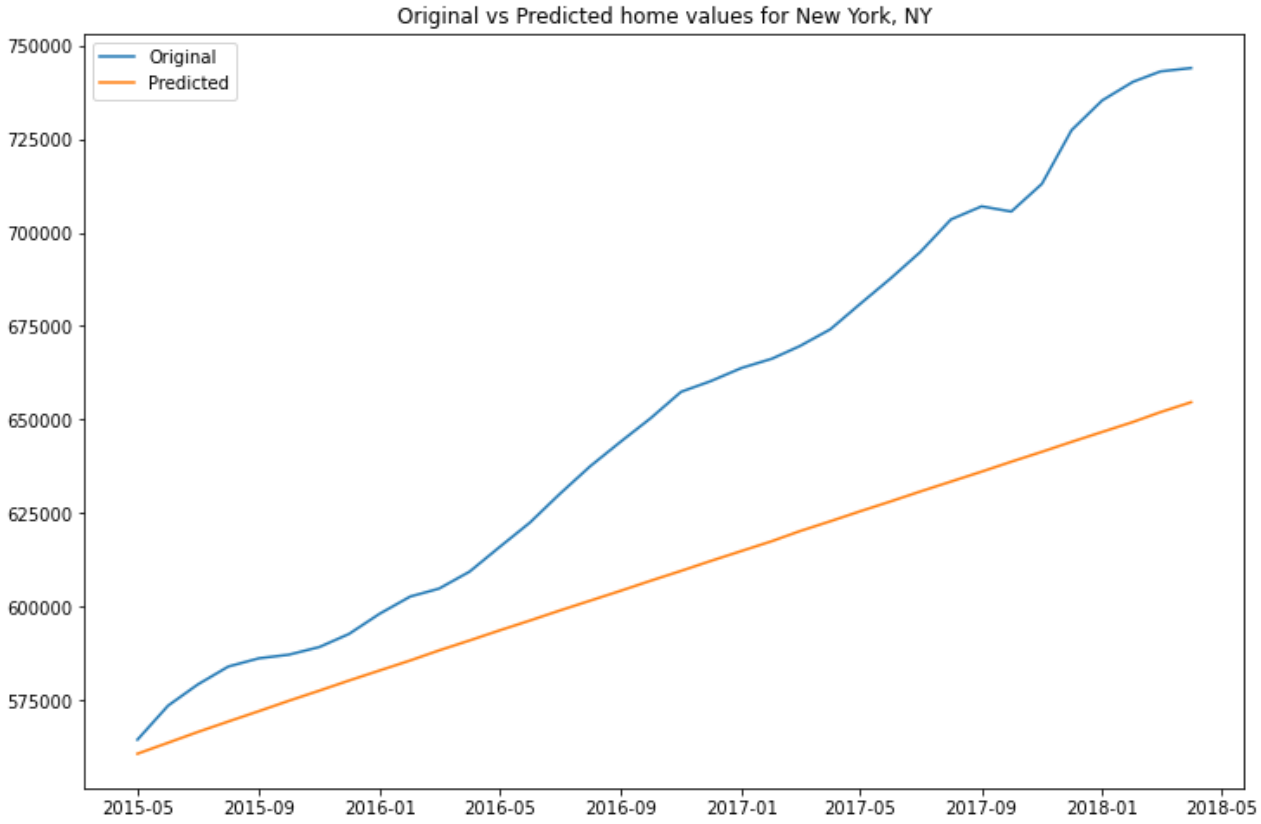
```
Sample:                     04-01-1996   HQIC                          3928.810
                          - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7822      5.496     -0.142      0.887     -11.553       9.989
ma.L1          0.7501      5.498      0.136      0.891     -10.026      11.526
ma.L2         -0.0381      0.171     -0.223      0.824      -0.373       0.297
ma.L3         -0.0118      0.064     -0.184      0.854      -0.138       0.115
sigma2       1.34e+06   1.82e+04     73.572      0.000     1.3e+06    1.38e+06
==============================================================================
===
Ljung-Box (L1) (Q):                  38.57   Jarque-Bera (JB):             7643
2.64
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              45.20   Skew:                            -
1.52
Prob(H) (two-sided):                  0.00   Kurtosis:                         9
2.84
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 5165.853590925947
                               SARIMAX Results
==============================================================================
Dep. Variable:        Los Angeles, CA   No. Observations:                  229
Model:                ARIMA(1, 2, 3)   Log Likelihood               -2075.658
Date:               Fri, 13 May 2022   AIC                           4161.315
Time:                       12:03:11   BIC                           4178.440
Sample:                   04-01-1996   HQIC                          4168.225
                        - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7735      0.205      3.774      0.000       0.372       1.175
ma.L1         -0.8365      0.209     -3.997      0.000      -1.247      -0.426
ma.L2         -0.0073      0.029     -0.253      0.800      -0.064       0.049
ma.L3         -0.0112      0.026     -0.422      0.673      -0.063       0.041
sigma2      5.032e+06   1.98e+05     25.411      0.000    4.64e+06    5.42e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   8.62   Jarque-Bera (JB):             1543
5.05
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              26.61   Skew:                            -
3.27
Prob(H) (two-sided):                  0.00   Kurtosis:                         4
2.86
==============================================================================
===

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

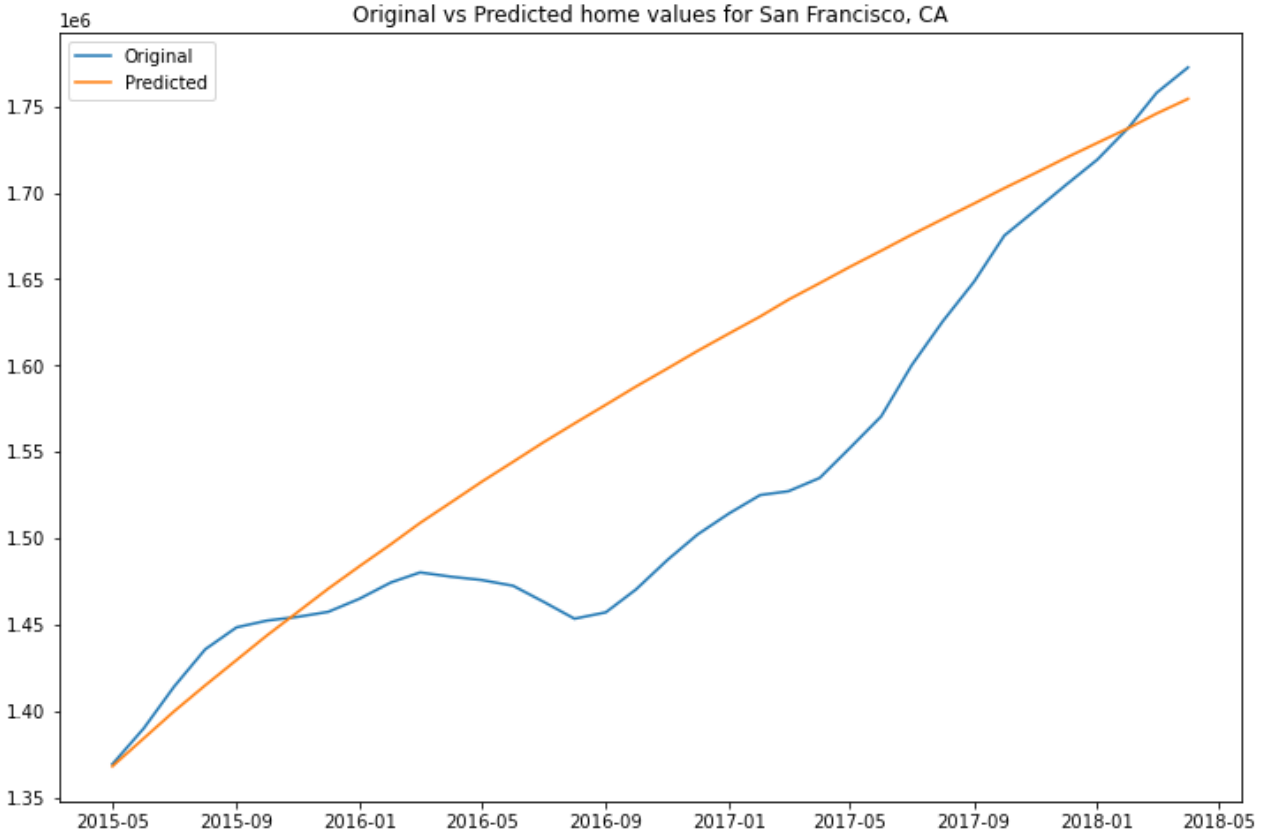
```
RMSE: 15666.858520860827
                               SARIMAX Results
==============================================================================
Dep. Variable:           San Jose, CA   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -2028.183
Date:                Fri, 13 May 2022   AIC                           4066.366
Time:                        12:03:11   BIC                           4083.490
Sample:                    04-01-1996   HQIC                          4073.276
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1906      8.385     -0.023      0.982     -16.624      16.243
ma.L1          0.1809      8.396      0.022      0.983     -16.276      16.638
ma.L2         -0.0174      0.069     -0.253      0.800      -0.152       0.118
ma.L3         -0.0074      0.112     -0.066      0.948      -0.227       0.213
sigma2      3.072e+06   1.37e+05     22.457      0.000     2.8e+06    3.34e+06
===================================================================================
===
Ljung-Box (L1) (Q):                   3.39   Jarque-Bera (JB):                79
3.12
Prob(Q):                              0.07   Prob(JB):
0.00
Heteroskedasticity (H):              15.62   Skew:
0.00
Prob(H) (two-sided):                  0.00   Kurtosis:                         1
2.16
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 33088.58251747072
                               SARIMAX Results
==============================================================================
Dep. Variable:             Chicago, IL   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -1863.466
Date:                Fri, 13 May 2022   AIC                           3736.932
Time:                        12:03:12   BIC                           3754.056
Sample:                    04-01-1996   HQIC                          3743.842
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9237      0.204      4.538      0.000       0.525       1.323
ma.L1         -0.9113      0.203     -4.493      0.000      -1.309      -0.514
ma.L2         -0.0235      0.023     -1.045      0.296      -0.068       0.021
ma.L3         -0.0037      0.027     -0.136      0.892      -0.056       0.049
sigma2      7.277e+05   2.76e+04     26.409      0.000    6.74e+05    7.82e+05
===================================================================================
===
Ljung-Box (L1) (Q):                   9.66   Jarque-Bera (JB):                79
5.02
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              26.99   Skew:
0.56
Prob(H) (two-sided):                  0.00   Kurtosis:                         1
2.10
===================================================================================
```

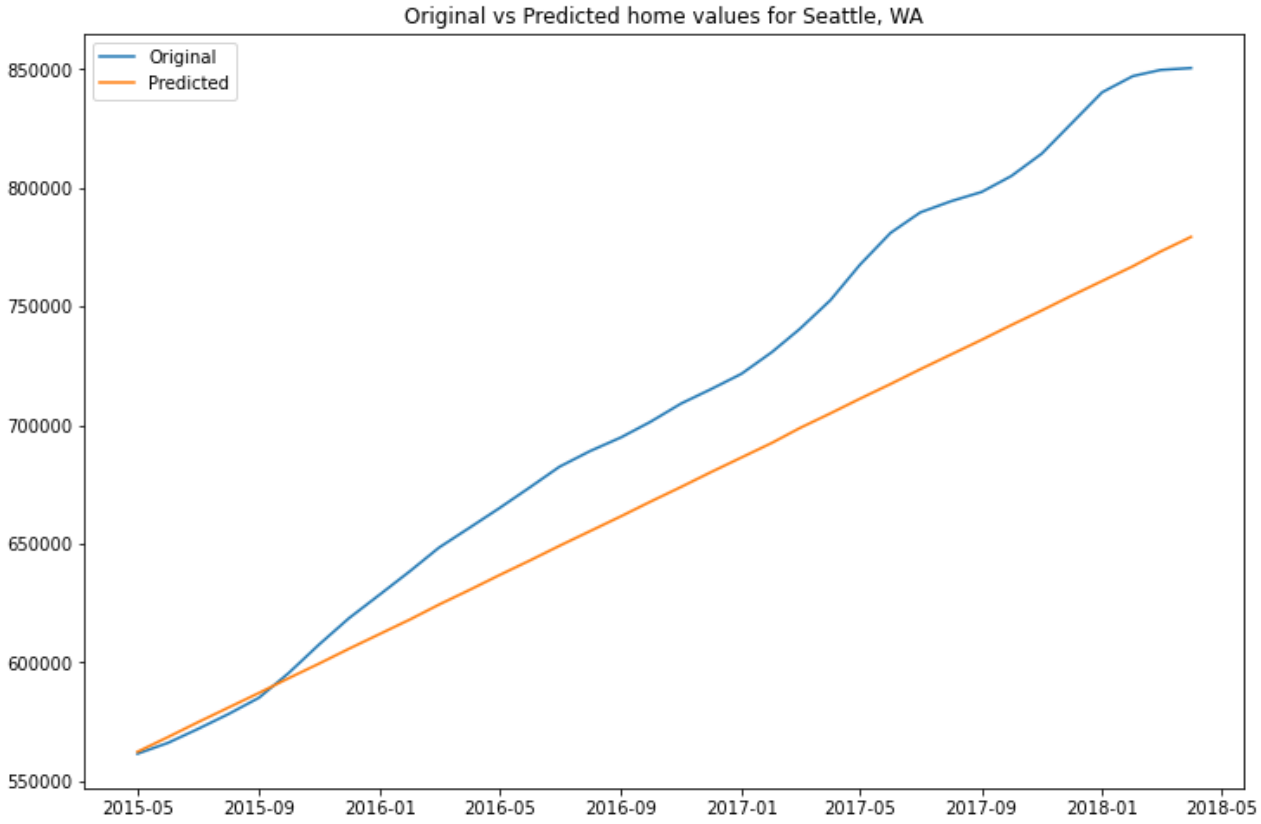
```
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 3961.5119734358436
                               SARIMAX Results
==============================================================================
Dep. Variable:          Baltimore, MD   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -1746.322
Date:                Fri, 13 May 2022   AIC                           3502.644
Time:                        12:03:12   BIC                           3519.768
Sample:                    04-01-1996   HQIC                          3509.554
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4565      0.464      0.984      0.325      -0.453       1.366
ma.L1         -0.4302      0.460     -0.935      0.350      -1.332       0.472
ma.L2         -0.1856      0.028     -6.664      0.000      -0.240      -0.131
ma.L3          0.0214      0.092      0.233      0.816      -0.159       0.202
sigma2      2.556e+05    1.2e+04     21.356      0.000    2.32e+05    2.79e+05
==============================================================================
===
Ljung-Box (L1) (Q):                   2.19   Jarque-Bera (JB):                49
0.25
Prob(Q):                              0.14   Prob(JB):
0.00
Heteroskedasticity (H):              13.02   Skew:
1.13
Prob(H) (two-sided):                  0.00   Kurtosis:
9.83
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 16300.327545138
                               SARIMAX Results
==============================================================================
Dep. Variable:              Boston, MA   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -1993.505
Date:                Fri, 13 May 2022   AIC                           3997.010
Time:                        12:03:13   BIC                           4014.134
Sample:                    04-01-1996   HQIC                          4003.920
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7180      0.135      5.335      0.000       0.454       0.982
ma.L1         -0.8022      0.136     -5.894      0.000      -1.069      -0.535
ma.L2         -0.0284      0.040     -0.714      0.475      -0.106       0.050
ma.L3         -0.0250      0.027     -0.935      0.350      -0.078       0.027
sigma2      2.273e+06   7.64e+04     29.768      0.000    2.12e+06    2.42e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   2.94   Jarque-Bera (JB):             1074
6.78
```

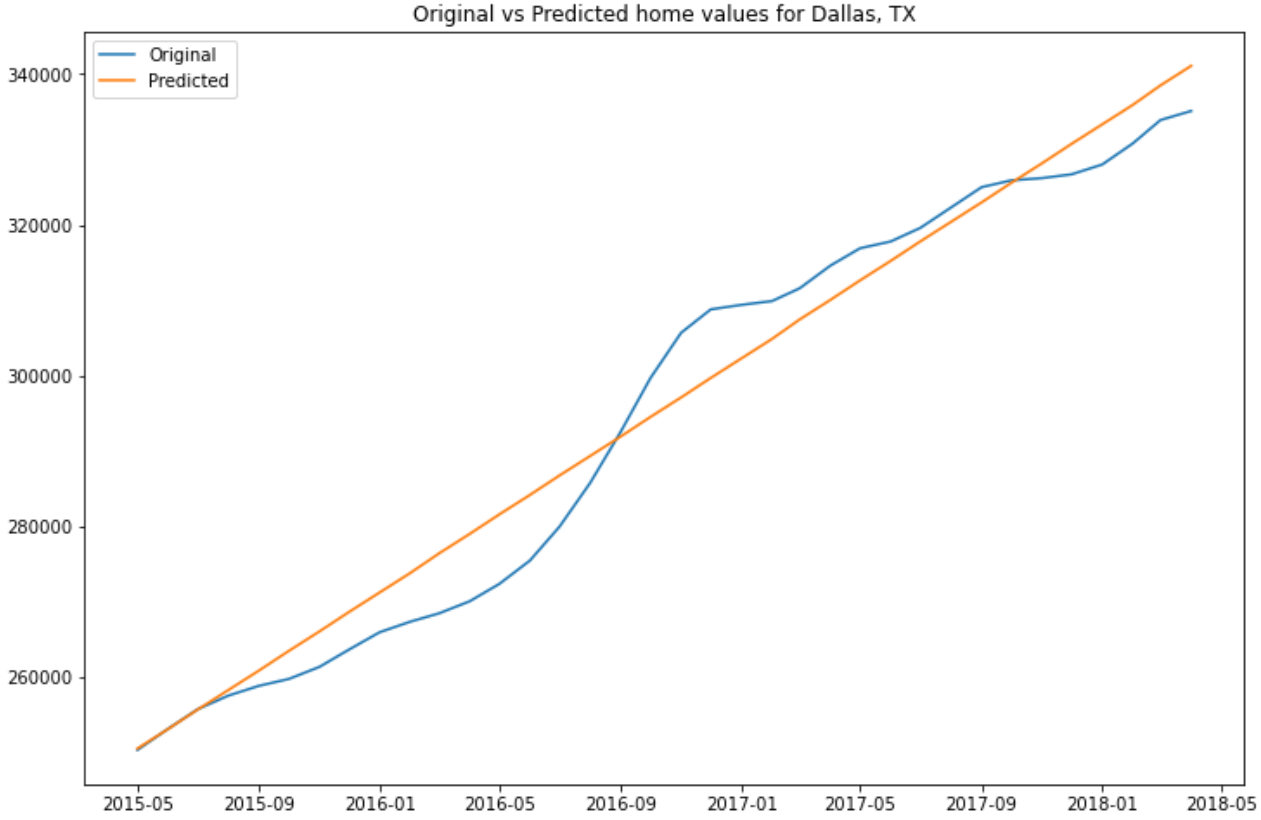
```
Prob(Q):                              0.09   Prob(JB):
0.00
Heteroskedasticity (H):              19.02   Skew:                            -
1.47
Prob(H) (two-sided):                  0.00   Kurtosis:                        3
6.58
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 53421.36806305324
```

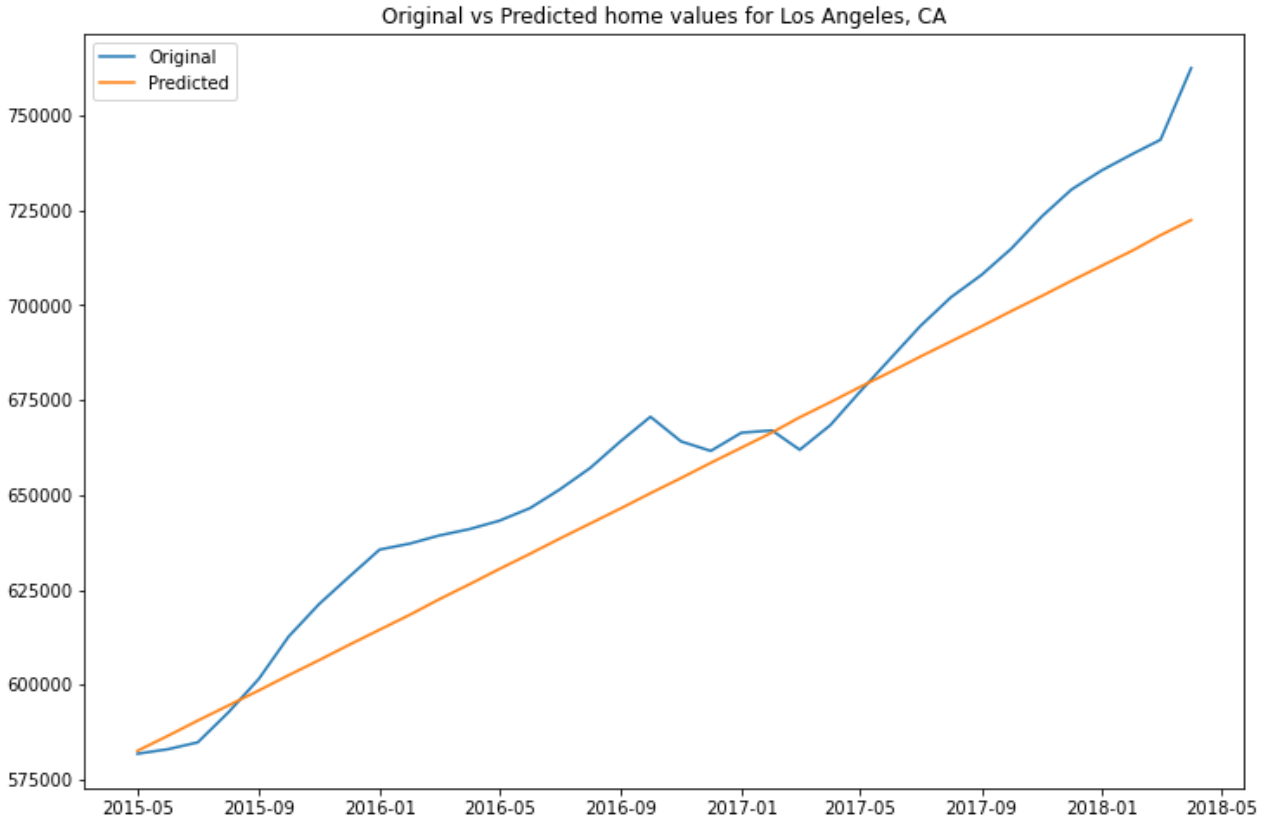


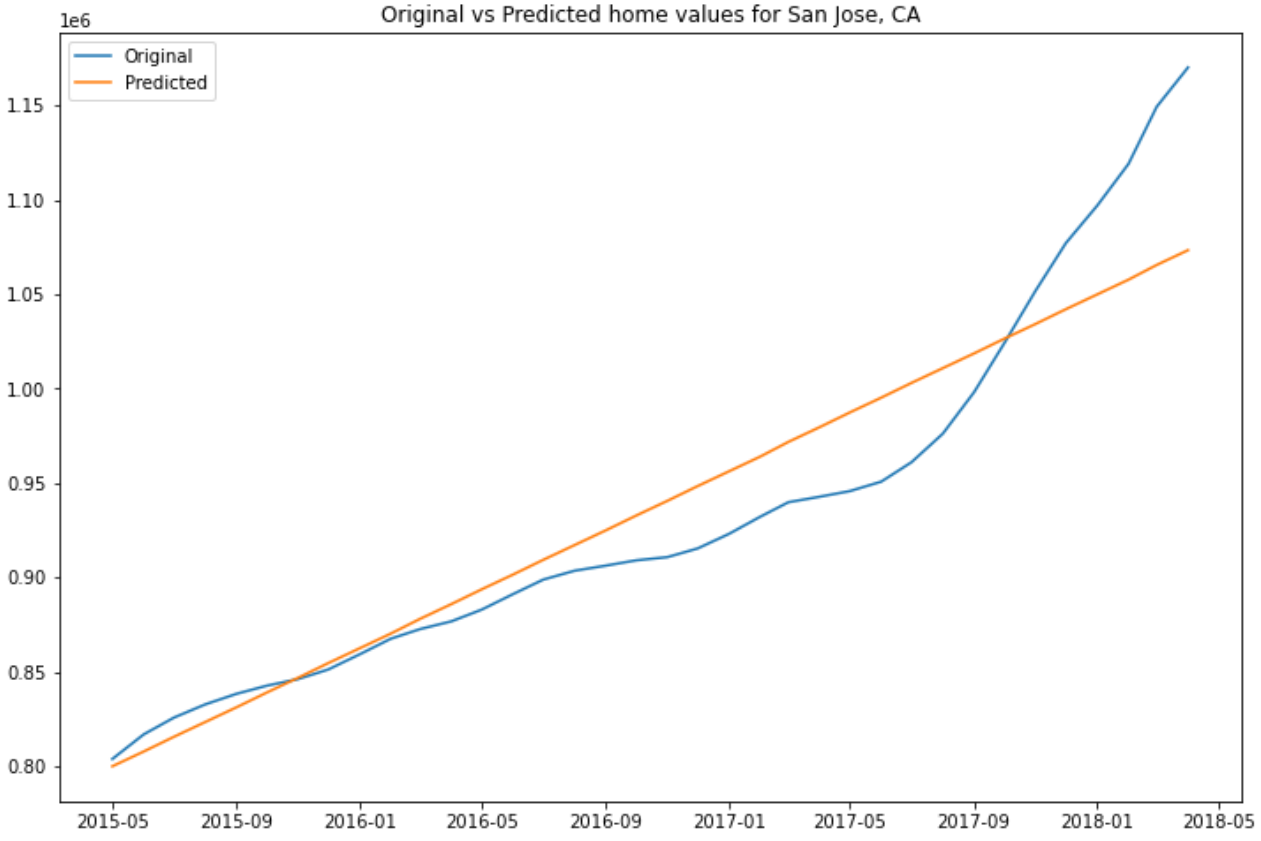

Original vs Predicted home values for New York, NY
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

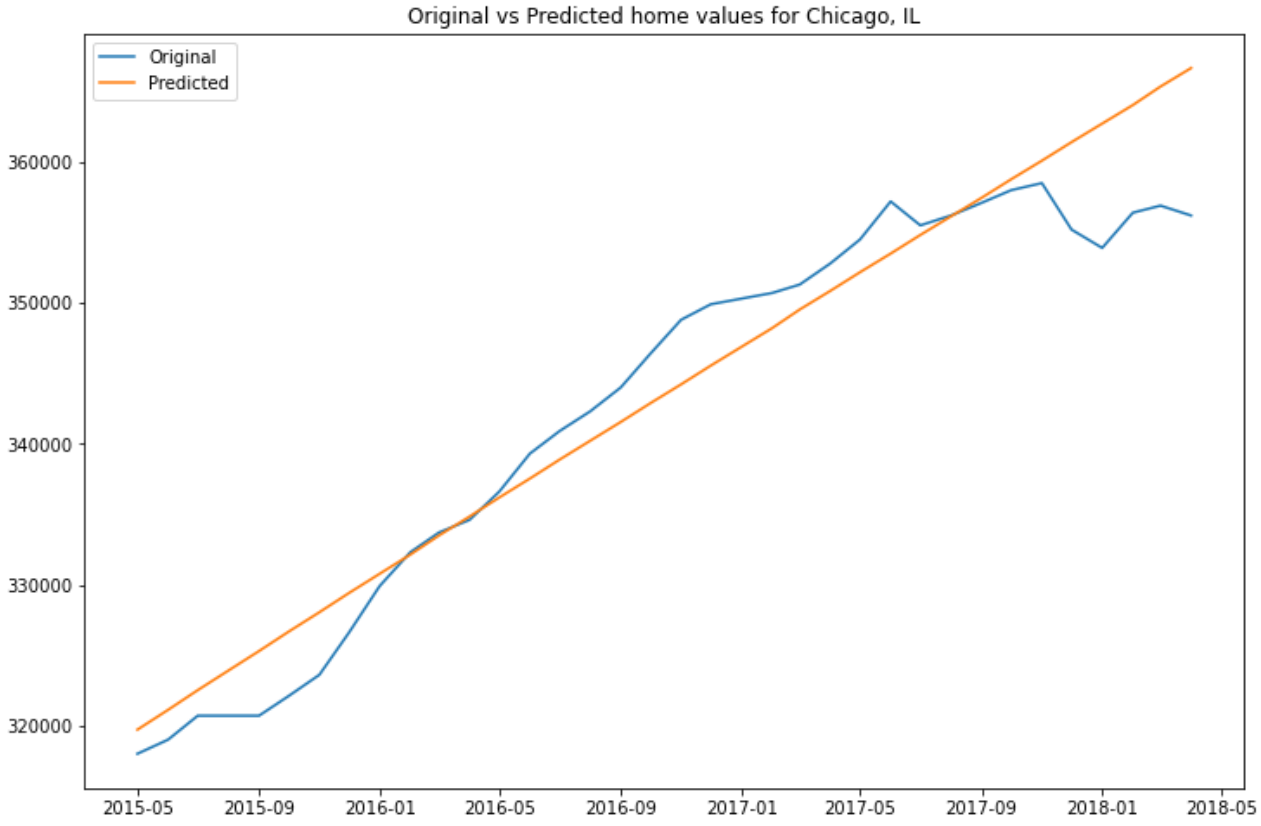

Original vs Predicted home values for San Francisco, CA
1e6
Original
Predicted
1.75
1.70
1.65
1.60
1.55
1.50
1.45
1.40
1.35
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

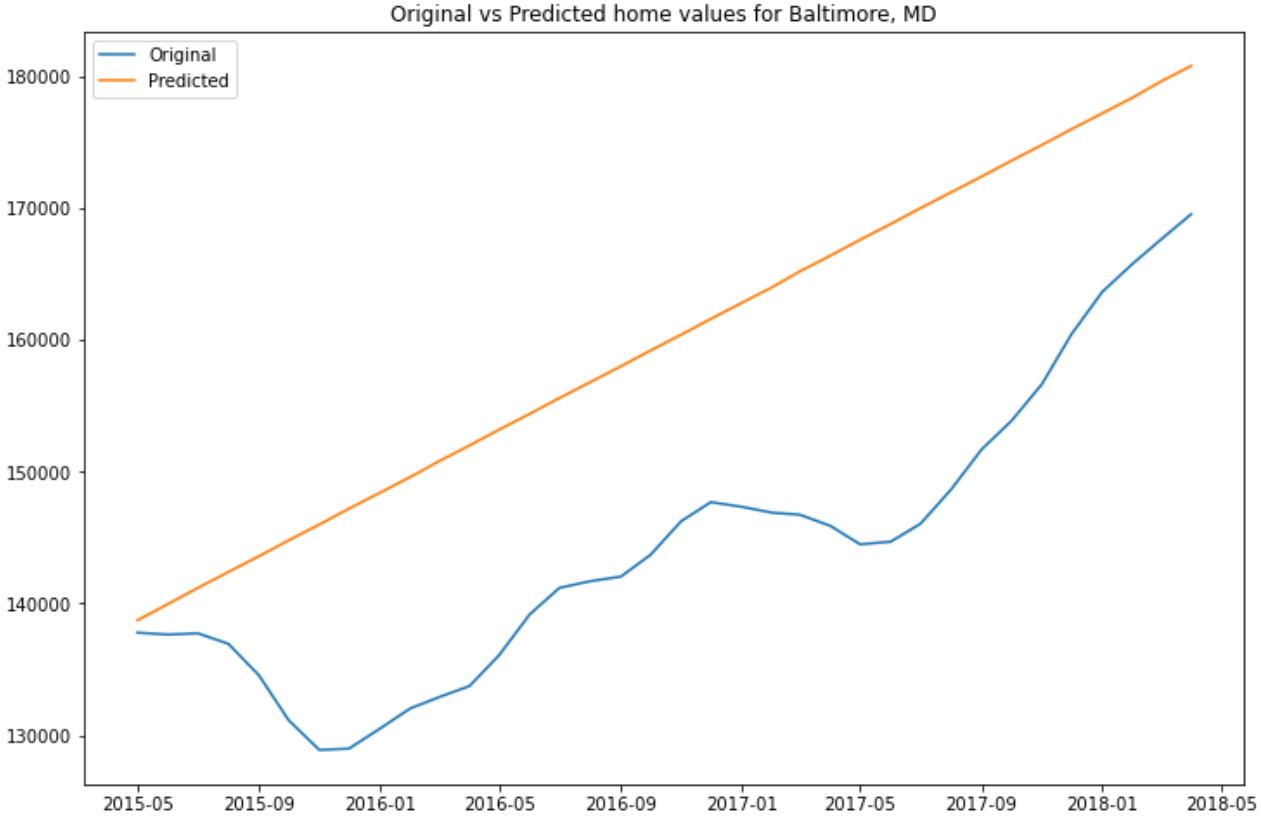

Original vs Predicted home values for Seattle, WA
Original
Predicted
850000
800000
750000
700000
650000
600000
550000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

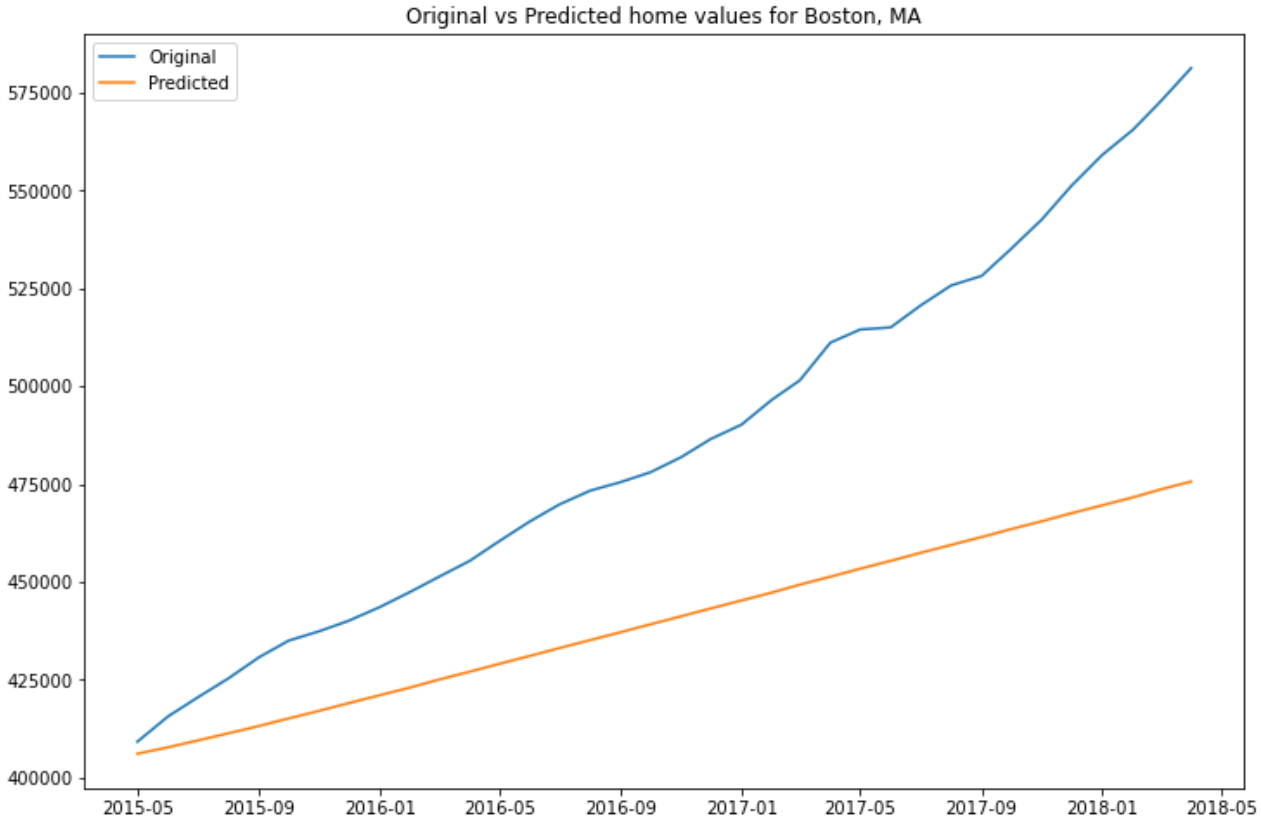

Original vs Predicted home values for Dallas, TX
Original
Predicted
340000
320000
300000
280000
260000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Los Angeles, CA
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for San Jose, CA
1e6
Original
Predicted
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Chicago, IL
Original
Predicted
360000
350000
340000
330000
320000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Baltimore, MD
Original
Predicted
180000
170000
160000
150000
140000
130000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05



In [78]:
```
# the best one yet
np.mean(rmse_list)
```

Out[78]: 32866.76222413008

In [79]:
```
#4 useful graphs
rmse_list = []
for city in city_list:
    city_model = arima_mod(city)
    city_model.model(train, test, 1,2,4)
    city_model.plot(test)
    rmse_list.append(city_model.rmse_)
```

SARIMAX Results

Dep. Variable:	Washington, DC	No. Observations:	229
Model:	ARIMA(1, 2, 4)	Log Likelihood	-1857.689
Date:	Fri, 13 May 2022	AIC	3727.379
Time:	12:03:17	BIC	3747.928
Sample:	04-01-1996	HQIC	3735.671
	- 04-01-2015		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9620	0.038	25.499	0.000	0.888	1.036
ma.L1	-0.9191	0.038	-24.057	0.000	-0.994	-0.844
ma.L2	-0.0343	0.018	-1.881	0.060	-0.070	0.001
ma.L3	-0.0248	0.037	-0.679	0.497	-0.096	0.047
ma.L4	-3.838e-05	0.040	-0.001	0.999	-0.079	0.079
sigma2	7.274e+05	4.42e+04	16.442	0.000	6.41e+05	8.14e+05

Ljung-Box (L1) (Q):	49.09	Jarque-Bera (JB):	10

```
1.76
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              15.97   Skew:
0.25
Prob(H) (two-sided):                  0.00   Kurtosis:
6.24
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 35463.97875215238
                               SARIMAX Results
================================================================================
Dep. Variable:           New York, NY   No. Observations:                  229
Model:                 ARIMA(1, 2, 4)   Log Likelihood               -2126.047
Date:                Fri, 13 May 2022   AIC                          4264.093
Time:                        12:03:18   BIC                          4284.643
Sample:                    04-01-1996   HQIC                         4272.385
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7026      0.114      6.149      0.000       0.479       0.927
ma.L1         -0.8372      0.128     -6.548      0.000      -1.088      -0.587
ma.L2         -0.0374      0.061     -0.618      0.537      -0.156       0.081
ma.L3         -0.0282      0.030     -0.934      0.350      -0.087       0.031
ma.L4         -0.0111      0.022     -0.510      0.610      -0.054       0.032
sigma2      7.465e+06   4.06e+05     18.394      0.000    6.67e+06    8.26e+06
================================================================================
===
Ljung-Box (L1) (Q):                   1.49   Jarque-Bera (JB):               27
5.26
Prob(Q):                              0.22   Prob(JB):
0.00
Heteroskedasticity (H):              13.52   Skew:                            -
0.23
Prob(H) (two-sided):                  0.00   Kurtosis:
8.37
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 7.7e+1
4. Standard errors may be unstable.
------------------------
------------------------
RMSE: 47635.41442610515
                               SARIMAX Results
================================================================================
Dep. Variable:      San Francisco, CA   No. Observations:                  229
Model:                 ARIMA(1, 2, 4)   Log Likelihood               -2143.193
Date:                Fri, 13 May 2022   AIC                          4298.386
Time:                        12:03:19   BIC                          4318.936
Sample:                    04-01-1996   HQIC                         4306.678
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
```

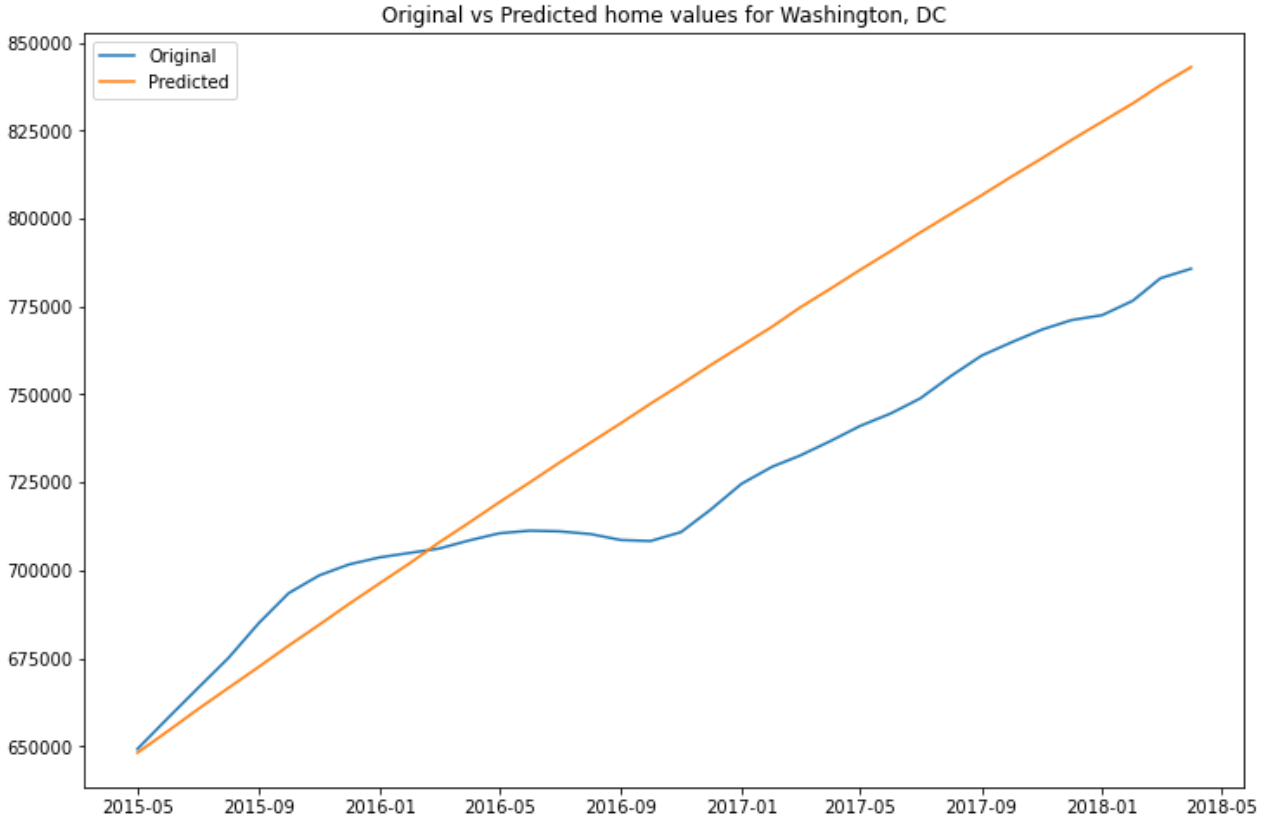
```
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9376      0.026     35.724      0.000       0.886       0.989
ma.L1         -0.9680      0.027    -35.675      0.000      -1.021      -0.915
ma.L2         -0.0153      0.011     -1.374      0.170      -0.037       0.007
ma.L3         -0.0072      0.028     -0.257      0.797      -0.062       0.048
ma.L4          0.0014      0.022      0.063      0.949      -0.042       0.045
sigma2      7.711e+06   3.77e-10   2.05e+16      0.000    7.71e+06    7.71e+06
===================================================================================
===
Ljung-Box (L1) (Q):                  21.21   Jarque-Bera (JB):                 9
4.29
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              19.00   Skew:
0.42
Prob(H) (two-sided):                  0.00   Kurtosis:
6.04
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 2.33e+
31. Standard errors may be unstable.
------------------------
------------------------
RMSE: 73801.33762172908
                               SARIMAX Results
==============================================================================
Dep. Variable:            Seattle, WA   No. Observations:                  229
Model:                 ARIMA(1, 2, 4)   Log Likelihood               -1882.557
Date:                Fri, 13 May 2022   AIC                           3777.113
Time:                        12:03:19   BIC                           3797.663
Sample:                    04-01-1996   HQIC                          3785.406
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0647     14.636     -0.004      0.996     -28.750      28.621
ma.L1          0.0751     14.637      0.005      0.996     -28.612      28.762
ma.L2         -0.0178      0.154     -0.116      0.908      -0.319       0.283
ma.L3         -0.0215      0.269     -0.080      0.936      -0.549       0.506
ma.L4          0.0010      0.289      0.003      0.997      -0.565       0.567
sigma2      6.772e+05   3.86e+04     17.562      0.000    6.02e+05    7.53e+05
===================================================================================
===
Ljung-Box (L1) (Q):                  24.07   Jarque-Bera (JB):                 2
4.49
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              23.51   Skew:
0.28
Prob(H) (two-sided):                  0.00   Kurtosis:
4.51
===================================================================================
===

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

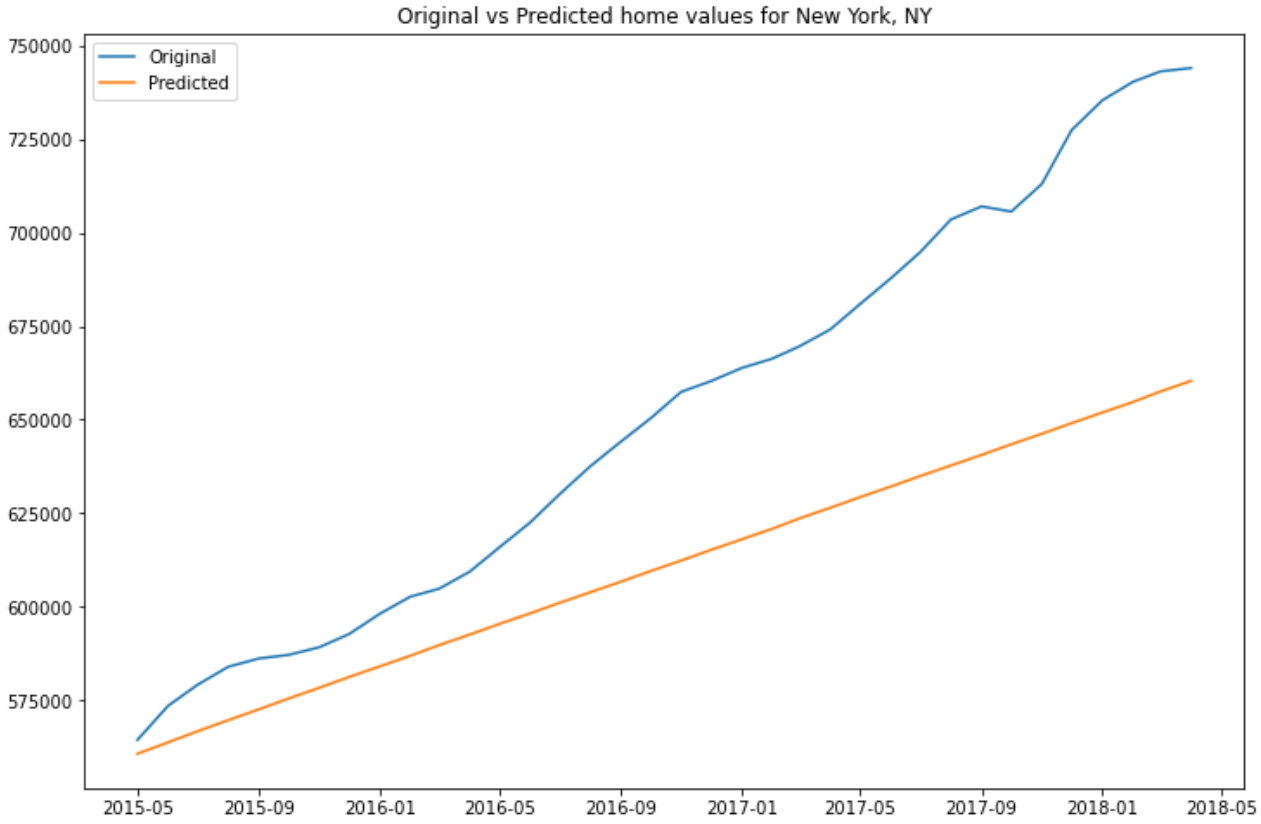
```
RMSE: 45447.80465875814
                               SARIMAX Results
==============================================================================
Dep. Variable:             Dallas, TX   No. Observations:                  229
Model:                 ARIMA(1, 2, 4)   Log Likelihood               -1955.773
Date:                Fri, 13 May 2022   AIC                           3923.547
Time:                        12:03:20   BIC                           3944.097
Sample:                    04-01-1996   HQIC                          3931.839
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.2928     18.771     -0.016      0.988     -37.084      36.498
ma.L1          0.2580     18.771      0.014      0.989     -36.532      37.048
ma.L2         -0.0280      0.652     -0.043      0.966      -1.306       1.250
ma.L3         -0.0157      0.326     -0.048      0.962      -0.654       0.623
ma.L4         -0.0061      0.176     -0.035      0.972      -0.351       0.339
sigma2       1.34e+06   1.76e+04     75.971      0.000    1.31e+06    1.37e+06
==============================================================================
===
Ljung-Box (L1) (Q):                  38.16   Jarque-Bera (JB):              7617
9.14
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              42.20   Skew:                            -
1.55
Prob(H) (two-sided):                  0.00   Kurtosis:                         9
2.69
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 5135.469034769434
                               SARIMAX Results
==============================================================================
Dep. Variable:          Los Angeles, CA   No. Observations:                  229
Model:                 ARIMA(1, 2, 4)   Log Likelihood               -2075.496
Date:                Fri, 13 May 2022   AIC                           4162.993
Time:                        12:03:21   BIC                           4183.543
Sample:                    04-01-1996   HQIC                          4171.285
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9393      0.035     26.574      0.000       0.870       1.009
ma.L1         -1.0028      0.028    -35.970      0.000      -1.057      -0.948
ma.L2          0.0041      0.022      0.185      0.853      -0.040       0.048
ma.L3         -0.0003      0.022     -0.013      0.990      -0.044       0.043
ma.L4          0.0083      0.025      0.334      0.738      -0.040       0.057
sigma2      5.044e+06   2.02e+05     25.013      0.000    4.65e+06    5.44e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   9.75   Jarque-Bera (JB):              1523
2.16
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              29.14   Skew:                            -
3.16
Prob(H) (two-sided):                  0.00   Kurtosis:                         4
```

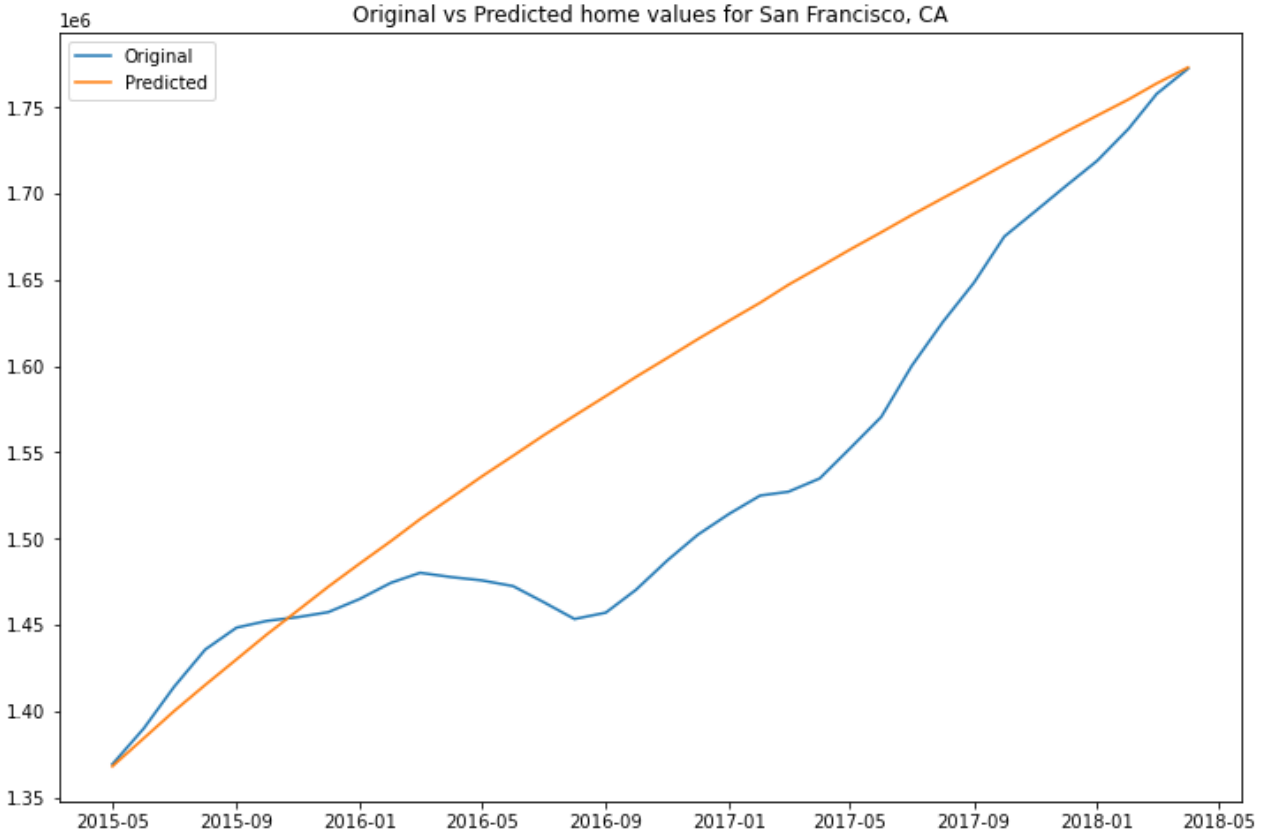
```
2.63
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 28505.966898725714
                               SARIMAX Results
==============================================================================
Dep. Variable:           San Jose, CA   No. Observations:                  229
Model:                 ARIMA(1, 2, 4)   Log Likelihood               -2028.336
Date:                Fri, 13 May 2022   AIC                           4068.672
Time:                        12:03:21   BIC                           4089.222
Sample:                    04-01-1996   HQIC                          4076.964
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1530     38.438     -0.004      0.997     -75.489      75.183
ma.L1          0.1432     38.438      0.004      0.997     -75.194      75.481
ma.L2         -0.0172      0.381     -0.045      0.964      -0.764       0.730
ma.L3         -0.0070      0.613     -0.011      0.991      -1.208       1.195
ma.L4       8.196e-05      0.189      0.000      1.000      -0.370       0.371
sigma2       3.03e+06   1.33e+05     22.792      0.000    2.77e+06    3.29e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   3.39   Jarque-Bera (JB):               79
2.83
Prob(Q):                              0.07   Prob(JB):
0.00
Heteroskedasticity (H):              15.61   Skew:
0.01
Prob(H) (two-sided):                  0.00   Kurtosis:                        1
2.16
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 33088.382702302835
                               SARIMAX Results
==============================================================================
Dep. Variable:             Chicago, IL   No. Observations:                  229
Model:                 ARIMA(1, 2, 4)   Log Likelihood               -1863.694
Date:                Fri, 13 May 2022   AIC                           3739.387
Time:                        12:03:21   BIC                           3759.937
Sample:                    04-01-1996   HQIC                          3747.679
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2543      8.177      0.031      0.975     -15.773      16.281
ma.L1         -0.2413      8.176     -0.030      0.976     -16.265      15.783
ma.L2         -0.0155      0.106     -0.145      0.884      -0.224       0.193
ma.L3         -0.0153      0.102     -0.149      0.882      -0.216       0.185
ma.L4          0.0010      0.151      0.007      0.995      -0.294       0.296
sigma2      7.373e+05    2.9e+04     25.449      0.000     6.8e+05    7.94e+05
```

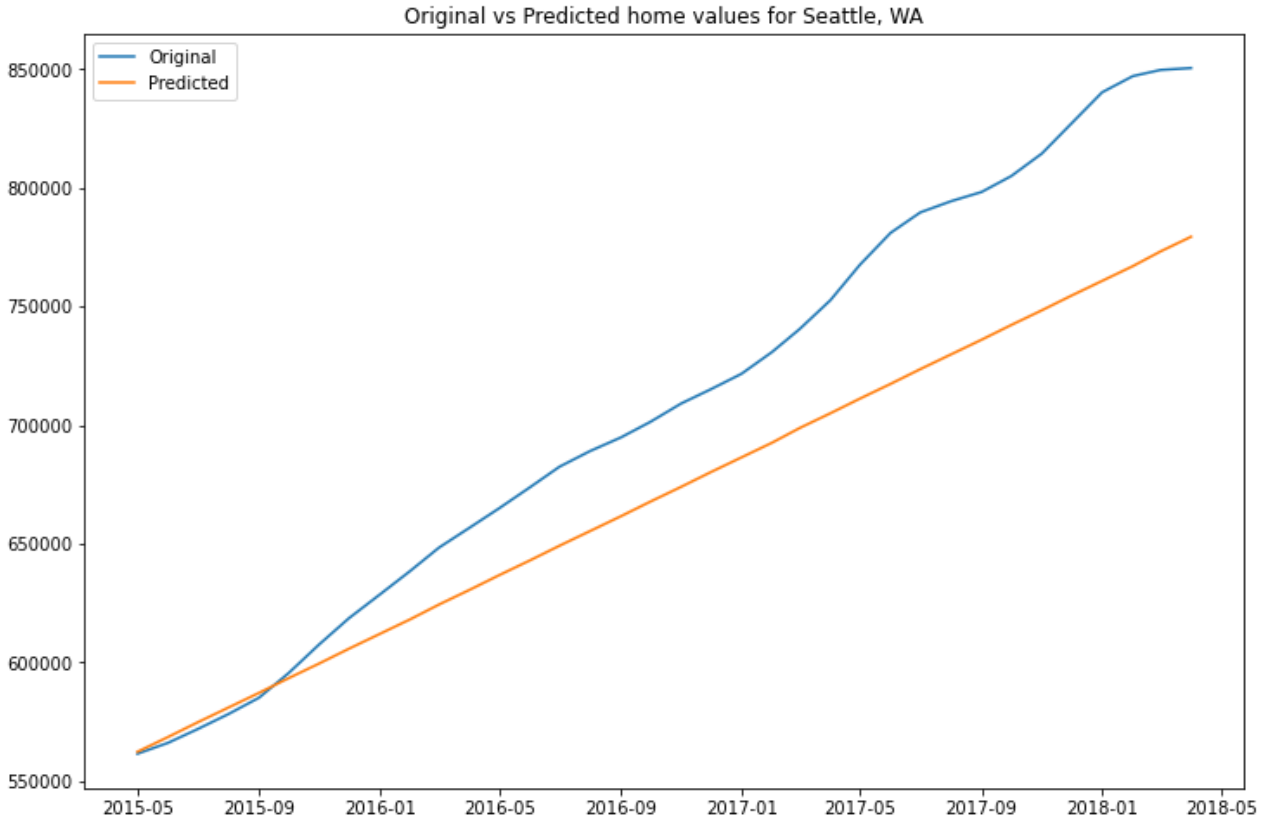
```
==========================================================================
===
Ljung-Box (L1) (Q):                  10.08   Jarque-Bera (JB):               77
1.32
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              28.10   Skew:
0.42
Prob(H) (two-sided):                  0.00   Kurtosis:                       1
1.99
==========================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
-----------------------
-----------------------
RMSE: 3978.8840289382492
                               SARIMAX Results
==========================================================================
Dep. Variable:          Baltimore, MD   No. Observations:                  229
Model:                ARIMA(1, 2, 4)   Log Likelihood               -1746.299
Date:               Fri, 13 May 2022   AIC                         3504.598
Time:                       12:03:22   BIC                         3525.148
Sample:                   04-01-1996   HQIC                        3512.890
                        - 04-01-2015
Covariance Type:                 opg
==========================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------
ar.L1          0.4160      1.301      0.320      0.749      -2.135       2.967
ma.L1         -0.3909      1.298     -0.301      0.763      -2.934       2.152
ma.L2         -0.1814      0.042     -4.366      0.000      -0.263      -0.100
ma.L3          0.0147      0.223      0.066      0.947      -0.422       0.451
ma.L4         -0.0052      0.092     -0.057      0.955      -0.186       0.175
sigma2      2.599e+05   1.25e+04     20.875      0.000    2.36e+05    2.84e+05
==========================================================================
===
Ljung-Box (L1) (Q):                   2.23   Jarque-Bera (JB):               49
0.90
Prob(Q):                              0.14   Prob(JB):
0.00
Heteroskedasticity (H):              13.12   Skew:
1.14
Prob(H) (two-sided):                  0.00   Kurtosis:
9.84
==========================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
-----------------------
-----------------------
RMSE: 16350.655040729132
                               SARIMAX Results
==========================================================================
Dep. Variable:              Boston, MA   No. Observations:                  229
Model:                ARIMA(1, 2, 4)   Log Likelihood               -1993.345
Date:               Fri, 13 May 2022   AIC                         3998.691
Time:                       12:03:22   BIC                         4019.241
Sample:                   04-01-1996   HQIC                        4006.983
                        - 04-01-2015
Covariance Type:                 opg
```

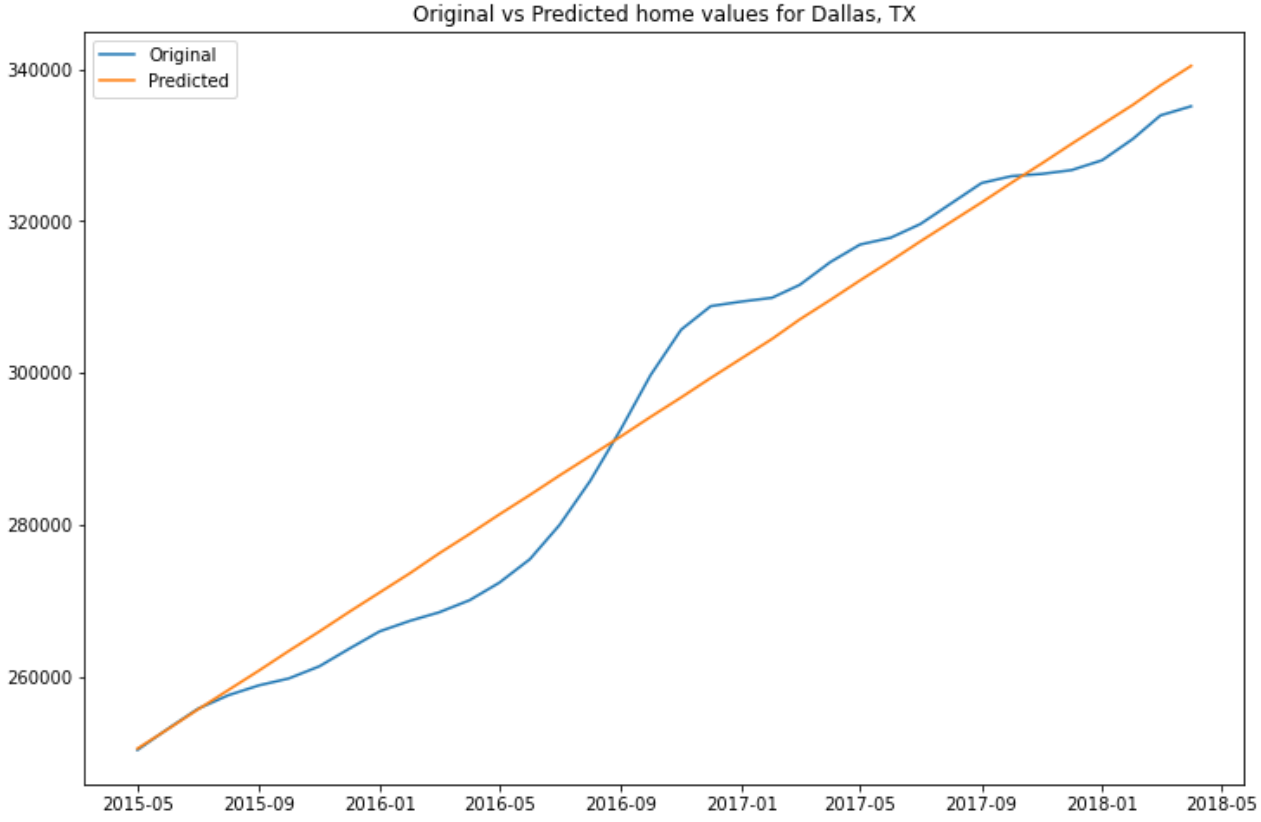
```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7385      0.205      3.602      0.000       0.337       1.140
ma.L1         -0.8228      0.198     -4.156      0.000      -1.211      -0.435
ma.L2         -0.0267      0.043     -0.629      0.529      -0.110       0.057
ma.L3         -0.0234      0.031     -0.753      0.452      -0.084       0.037
ma.L4          0.0037      0.025      0.148      0.882      -0.046       0.053
sigma2      2.309e+06   7.93e+04     29.121      0.000    2.15e+06    2.46e+06
==============================================================================
===
Ljung-Box (L1) (Q):                   2.94   Jarque-Bera (JB):             1070
1.06
Prob(Q):                              0.09   Prob(JB):
0.00
Heteroskedasticity (H):              18.98   Skew:                           -
1.46
Prob(H) (two-sided):                  0.00   Kurtosis:                        3
6.51
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 53103.38071071858
```

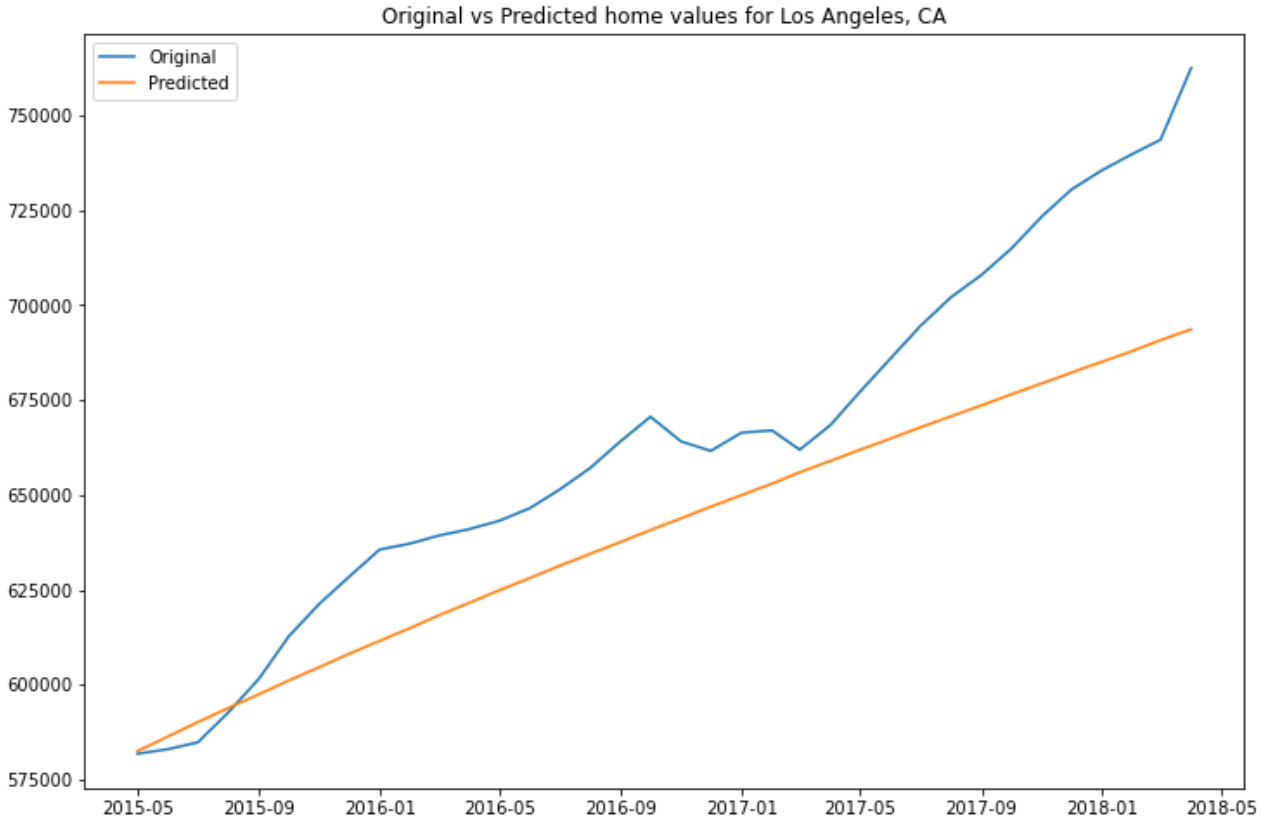


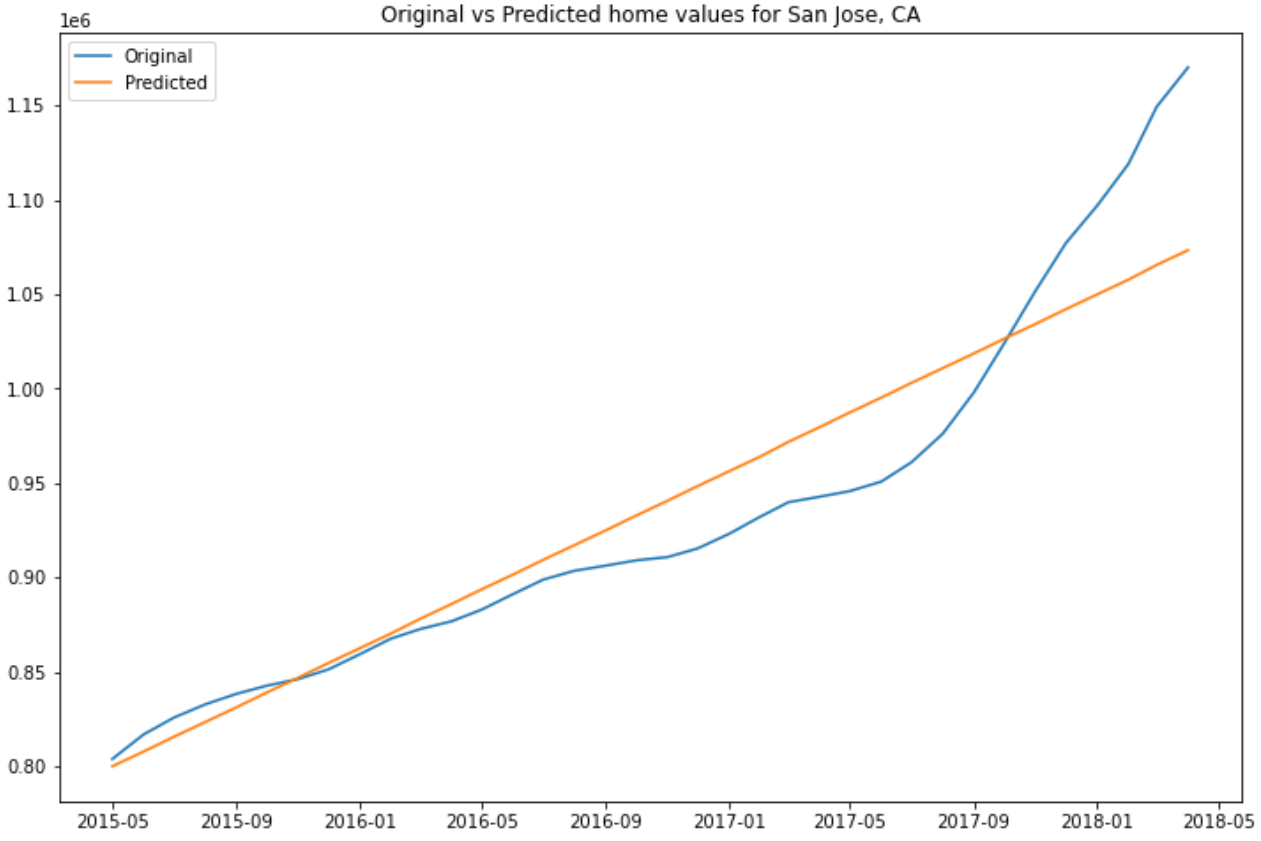

Original vs Predicted home values for New York, NY
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

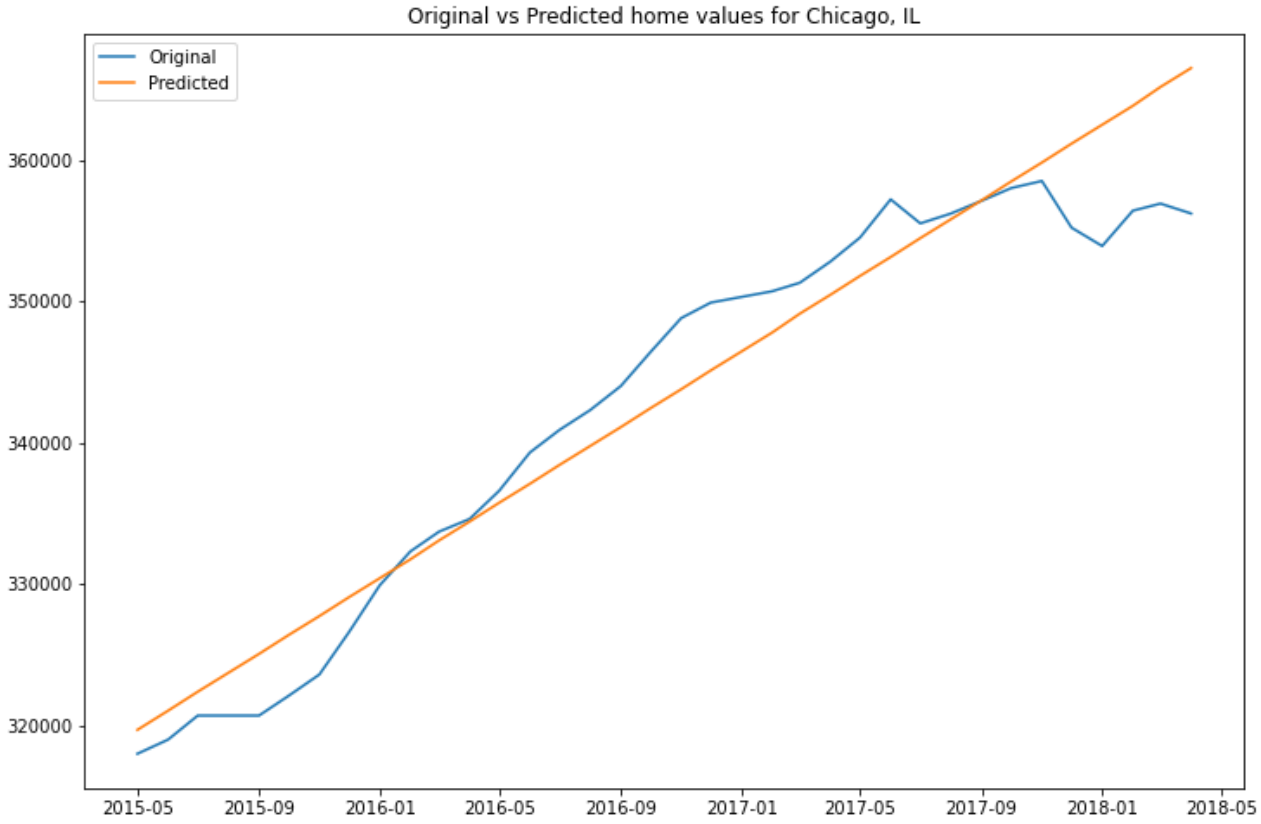

Original vs Predicted home values for San Francisco, CA
1e6
Original
Predicted
1.75
1.70
1.65
1.60
1.55
1.50
1.45
1.40
1.35
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

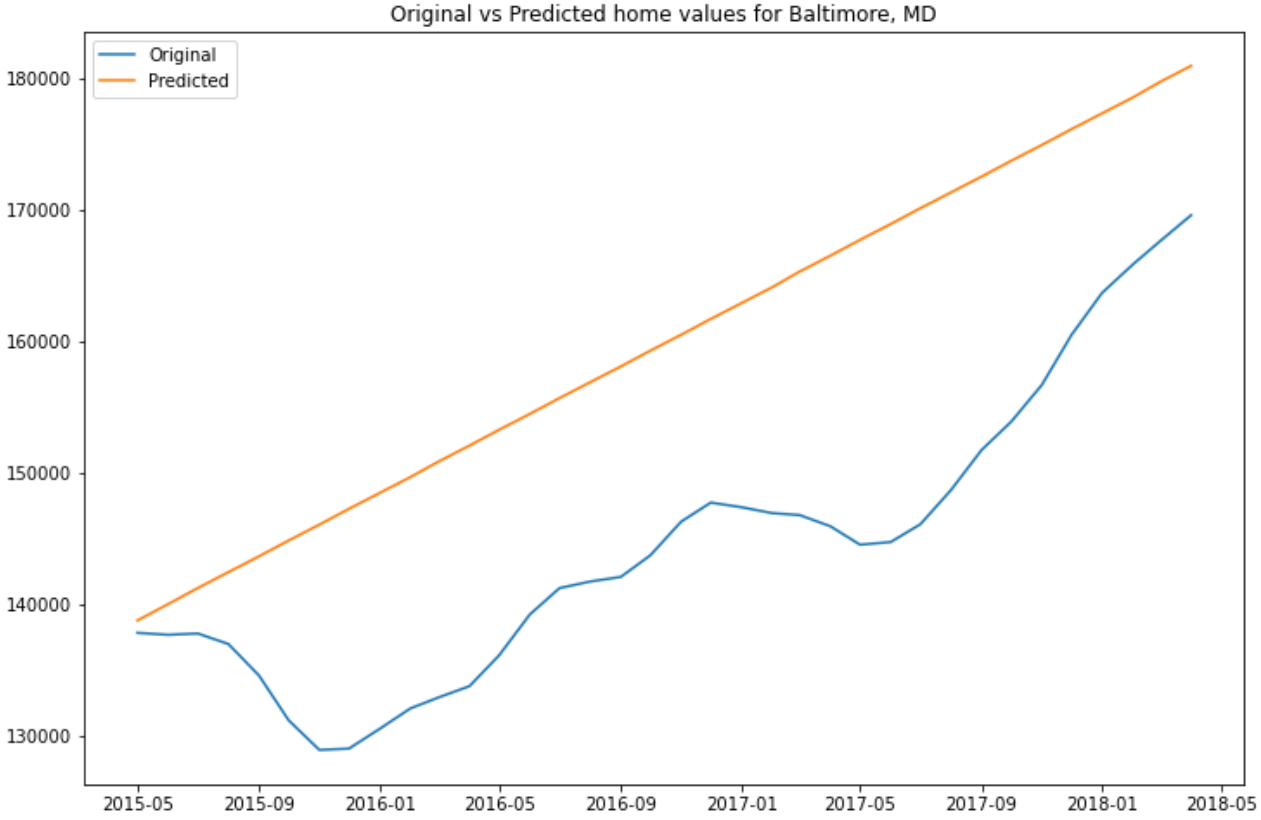

Original vs Predicted home values for Seattle, WA
Original
Predicted
850000
800000
750000
700000
650000
600000
550000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

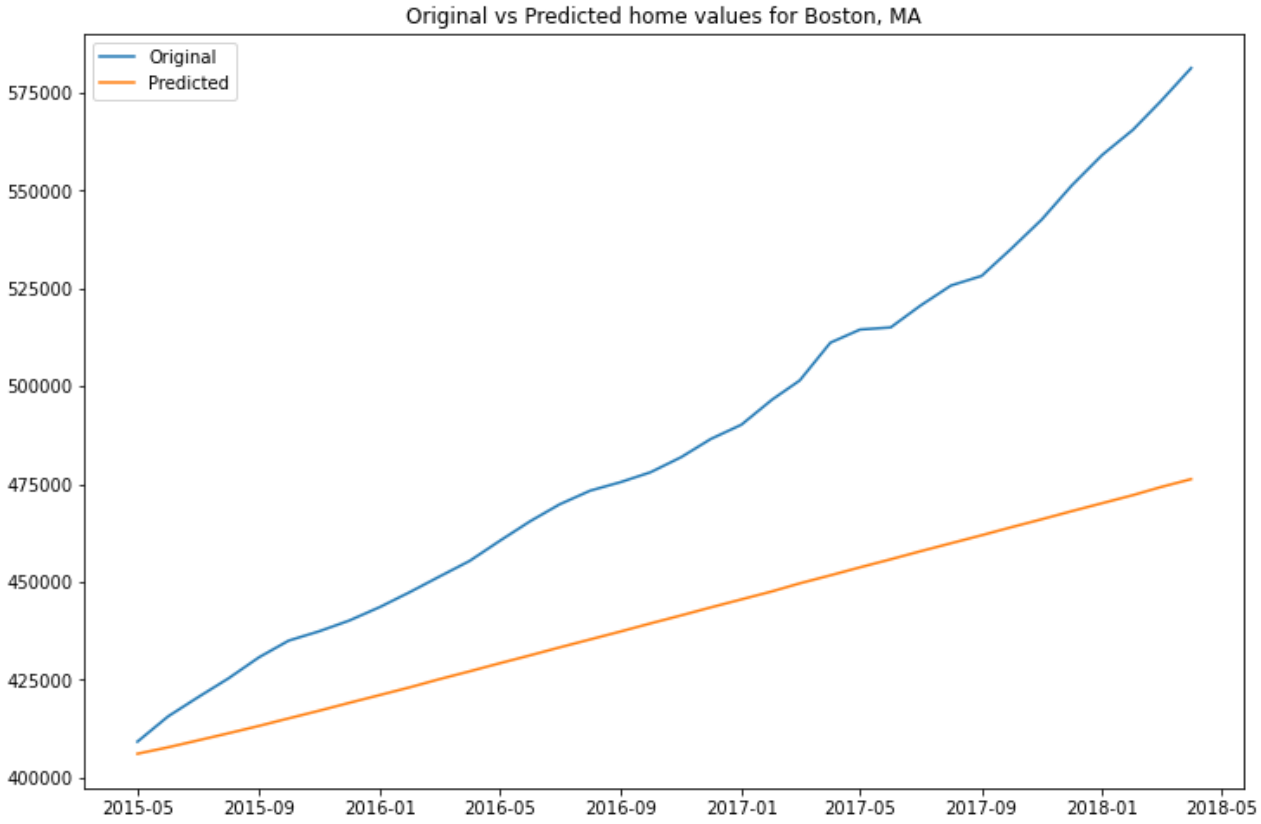

Original vs Predicted home values for Dallas, TX
Original
Predicted
340000
320000
300000
280000
260000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Los Angeles, CA
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for San Jose, CA
1e6
Original
Predicted
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Chicago, IL
Original
Predicted
360000
350000
340000
330000
320000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Baltimore, MD
Original
Predicted
180000
170000
160000
150000
140000
130000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05



In [80]:
```
# moving in the wrong direction again
np.mean(rmse_list)
```

Out[80]:
```
34251.127387492874
```

In [81]:
```
# try a different balance
#rmse_list = []
#for city in city_list:
#    city_model = arima_mod(city)
#    city_model.model(train, test, 3,1,3)
#    city_model.plot(test)
#    rmse_list.append(city_model.rmse_)
```

In [82]:
```
# couldn't converge on an answer for the fourth city
# np.mean(rmse_list)
```

In [90]:
```
# another try for a different p parameter
# re-run the cell in case of an error. There might be some randomness in determi
rmse_list = []
for city in city_list:
    city_model = arima_mod(city)
    city_model.model(train, test, 4,1,2)
    city_model.plot(test)
    rmse_list.append(city_model.rmse_)
```

```
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:966: UserWarning:

Non-stationary starting autoregressive parameters found. Using zeros as starting
parameters.

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
```

```
pace/sarimax.py:978: UserWarning:

Non-invertible starting MA parameters found. Using zeros as starting parameters.

                               SARIMAX Results
==========================================================================
Dep. Variable:         Washington, DC   No. Observations:                  229
Model:                 ARIMA(4, 1, 2)   Log Likelihood               -2173.633
Date:                Fri, 13 May 2022   AIC                           4361.265
Time:                        12:06:15   BIC                           4385.270
Sample:                    04-01-1996   HQIC                          4370.951
                         - 04-01-2015
Covariance Type:                  opg
==========================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------
ar.L1          2.0876      0.950      2.197      0.028       0.225       3.950
ar.L2         -1.0685      1.041     -1.027      0.305      -3.108       0.971
ar.L3         -0.1373      0.018     -7.572      0.000      -0.173      -0.102
ar.L4          0.1181      0.111      1.067      0.286      -0.099       0.335
ma.L1         -1.9646      0.948     -2.072      0.038      -3.823      -0.106
ma.L2          0.9648      0.922      1.047      0.295      -0.842       2.771
sigma2      6.941e+06      1e-07   6.94e+13      0.000    6.94e+06    6.94e+06
==========================================================================
===
Ljung-Box (L1) (Q):                 158.01   Jarque-Bera (JB):              138
4.48
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               0.62   Skew:                           -
1.81
Prob(H) (two-sided):                  0.04   Kurtosis:                        1
4.52
==========================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 1.1e+3
1. Standard errors may be unstable.
------------------------
------------------------
RMSE: 36497.81832203444
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:966: UserWarning:

Non-stationary starting autoregressive parameters found. Using zeros as starting
parameters.

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:978: UserWarning:

Non-invertible starting MA parameters found. Using zeros as starting parameters.

                               SARIMAX Results
==========================================================================
Dep. Variable:           New York, NY   No. Observations:                  229
Model:                 ARIMA(4, 1, 2)   Log Likelihood               -2228.439
Date:                Fri, 13 May 2022   AIC                           4470.877
Time:                        12:06:16   BIC                           4494.883
Sample:                    04-01-1996   HQIC                          4480.563
                         - 04-01-2015
Covariance Type:                  opg
==========================================================================
```

```
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1412      1.186      0.119      0.905      -2.183       2.465
ar.L2          0.8106      1.176      0.690      0.490      -1.493       3.115
ar.L3         -0.0264      0.028     -0.950      0.342      -0.081       0.028
ar.L4         -0.0008      0.012     -0.063      0.949      -0.025       0.023
ma.L1         -0.0808      1.186     -0.068      0.946      -2.405       2.244
ma.L2         -0.7844      1.104     -0.711      0.477      -2.948       1.379
sigma2      6.917e+06   9.46e-07   7.31e+12      0.000    6.92e+06    6.92e+06
===================================================================================
===
Ljung-Box (L1) (Q):                  92.66   Jarque-Bera (JB):                 6
1.21
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               3.03   Skew:                            -
0.13
Prob(H) (two-sided):                  0.00   Kurtosis:
5.53
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 8.63e+
28. Standard errors may be unstable.
------------------------
------------------------
RMSE: 93522.44542382043
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:966: UserWarning:

Non-stationary starting autoregressive parameters found. Using zeros as starting
parameters.

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:978: UserWarning:

Non-invertible starting MA parameters found. Using zeros as starting parameters.

                               SARIMAX Results
==============================================================================
Dep. Variable:      San Francisco, CA   No. Observations:                  229
Model:                 ARIMA(4, 1, 2)   Log Likelihood               -2082.341
Date:                Fri, 13 May 2022   AIC                           4178.682
Time:                        12:06:17   BIC                           4202.688
Sample:                    04-01-1996   HQIC                          4188.368
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.4416   3.39e-08   4.25e+07      0.000       1.442       1.442
ar.L2         -0.4678   1.35e-08  -3.45e+07      0.000      -0.468      -0.468
ar.L3          0.6108   6.67e-09   9.15e+07      0.000       0.611       0.611
ar.L4         -0.5846   2.65e-10  -2.21e+09      0.000      -0.585      -0.585
ma.L1         -2.0000   7.78e-09  -2.57e+08      0.000      -2.000      -2.000
ma.L2          1.0000   2.56e-08   3.91e+07      0.000       1.000       1.000
sigma2      9.747e+08   2.94e-17   3.32e+25      0.000    9.75e+08    9.75e+08
===================================================================================
===
Ljung-Box (L1) (Q):                  55.95   Jarque-Bera (JB):             12520
9.66
Prob(Q):                              0.00   Prob(JB):
```

```
0.00
Heteroskedasticity (H):               0.00   Skew:
9.76
Prob(H) (two-sided):                  0.00   Kurtosis:                        11
6.13
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 5.74e+
33. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 191776021.24846232
                               SARIMAX Results
================================================================================
Dep. Variable:            Seattle, WA   No. Observations:                   229
Model:                 ARIMA(4, 1, 2)   Log Likelihood                -2136.601
Date:                Fri, 13 May 2022   AIC                             4287.202
Time:                        12:06:18   BIC                             4311.208
Sample:                    04-01-1996   HQIC                            4296.888
                         - 04-01-2015
Covariance Type:                 opg
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
ar.L1          0.0615      0.070      0.876      0.381      -0.076       0.199
ar.L2          1.0522      0.076     13.919      0.000       0.904       1.200
ar.L3         -0.0556      0.010     -5.512      0.000      -0.075      -0.036
ar.L4         -0.0585      0.007     -7.824      0.000      -0.073      -0.044
ma.L1         -0.0036      0.070     -0.051      0.959      -0.142       0.135
ma.L2         -0.9929      0.072    -13.880      0.000      -1.133      -0.853
sigma2      7.636e+06   1.55e-09   4.91e+15      0.000    7.64e+06    7.64e+06
================================================================================
===
Ljung-Box (L1) (Q):                 182.45   Jarque-Bera (JB):                5
1.13
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               3.73   Skew:                           -
1.00
Prob(H) (two-sided):                  0.00   Kurtosis:
4.17
================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 7.94e+
31. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 143290.6756209475
                               SARIMAX Results
================================================================================
Dep. Variable:              Dallas, TX   No. Observations:                   229
Model:                 ARIMA(4, 1, 2)   Log Likelihood                -1938.582
Date:                Fri, 13 May 2022   AIC                             3891.164
Time:                        12:06:18   BIC                             3915.170
Sample:                    04-01-1996   HQIC                            3900.850
                         - 04-01-2015
Covariance Type:                 opg
```

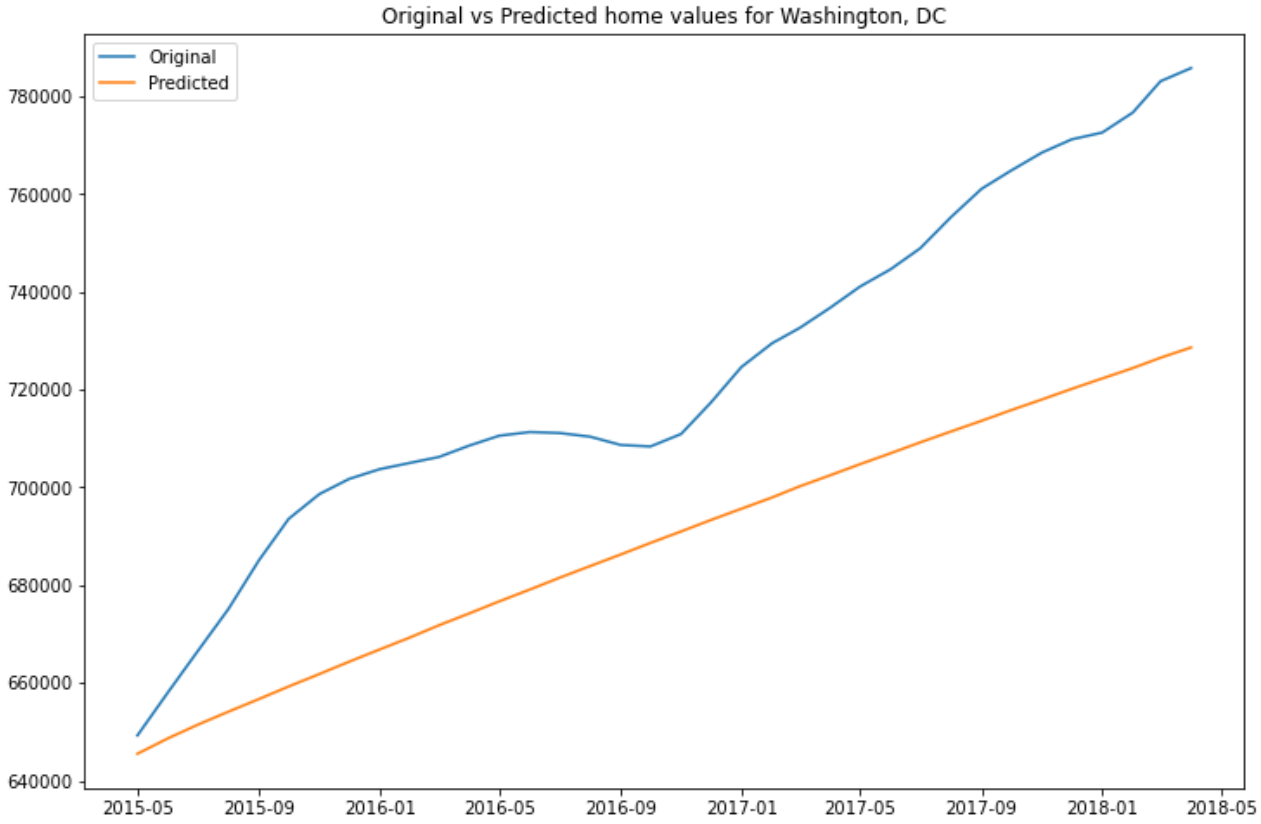
```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.9040      0.012    152.530      0.000       1.880       1.928
ar.L2         -1.0208      0.024    -41.989      0.000      -1.068      -0.973
ar.L3         -0.0071      0.026     -0.269      0.788      -0.059       0.045
ar.L4          0.0135      0.013      1.043      0.297      -0.012       0.039
ma.L1         -1.8852      0.024    -79.288      0.000      -1.932      -1.839
ma.L2          0.9997      0.025     40.437      0.000       0.951       1.048
sigma2      1.242e+06   3.48e-08   3.57e+13      0.000    1.24e+06    1.24e+06
==============================================================================
===
Ljung-Box (L1) (Q):                  21.84   Jarque-Bera (JB):              4662
7.33
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              13.17   Skew:                            -
5.99
Prob(H) (two-sided):                  0.00   Kurtosis:                         7
2.02
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 2.9e+2
8. Standard errors may be unstable.
------------------------
------------------------
RMSE: 55469.6712946405
                               SARIMAX Results
==============================================================================
Dep. Variable:        Los Angeles, CA   No. Observations:                  229
Model:                ARIMA(4, 1, 2)   Log Likelihood               -2313.983
Date:               Fri, 13 May 2022   AIC                           4641.967
Time:                       12:06:19   BIC                           4665.972
Sample:                   04-01-1996   HQIC                          4651.652
                        - 04-01-2015
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          2.0035      0.009    213.286      0.000       1.985       2.022
ar.L2         -1.0102      0.011    -90.199      0.000      -1.032      -0.988
ar.L3         -0.0391      0.007     -5.546      0.000      -0.053      -0.025
ar.L4          0.0426      0.005      9.031      0.000       0.033       0.052
ma.L1         -1.9205      0.009   -204.917      0.000      -1.939      -1.902
ma.L2          0.9258      0.009     99.649      0.000       0.908       0.944
sigma2      5.041e+06   5.86e-10    8.6e+15      0.000    5.04e+06    5.04e+06
==============================================================================
===
Ljung-Box (L1) (Q):                 136.48   Jarque-Bera (JB):                5
4.86
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.17   Skew:                            -
0.52
Prob(H) (two-sided):                  0.50   Kurtosis:
5.16
==============================================================================
===

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

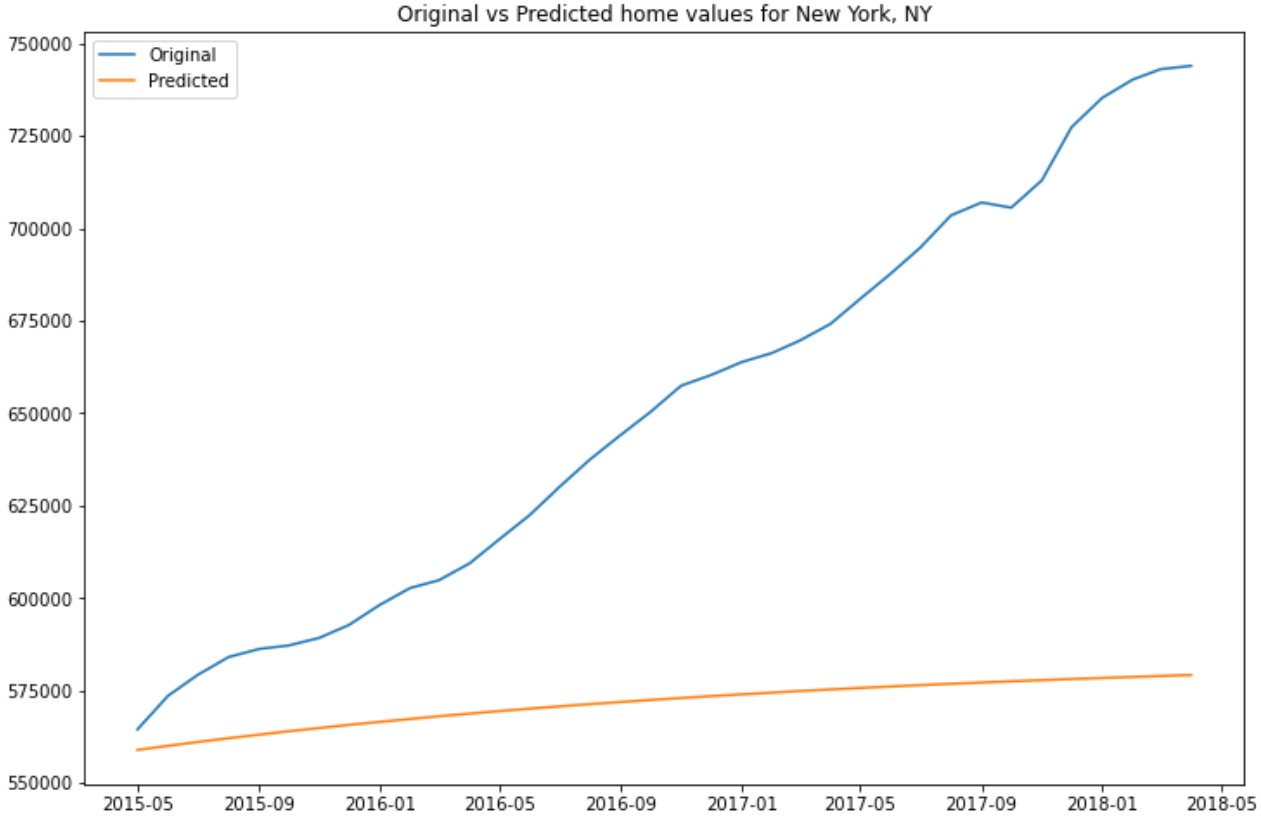
```
tep).
[2] Covariance matrix is singular or near-singular, with condition number 2.28e+
31. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 83170.50502573523
```

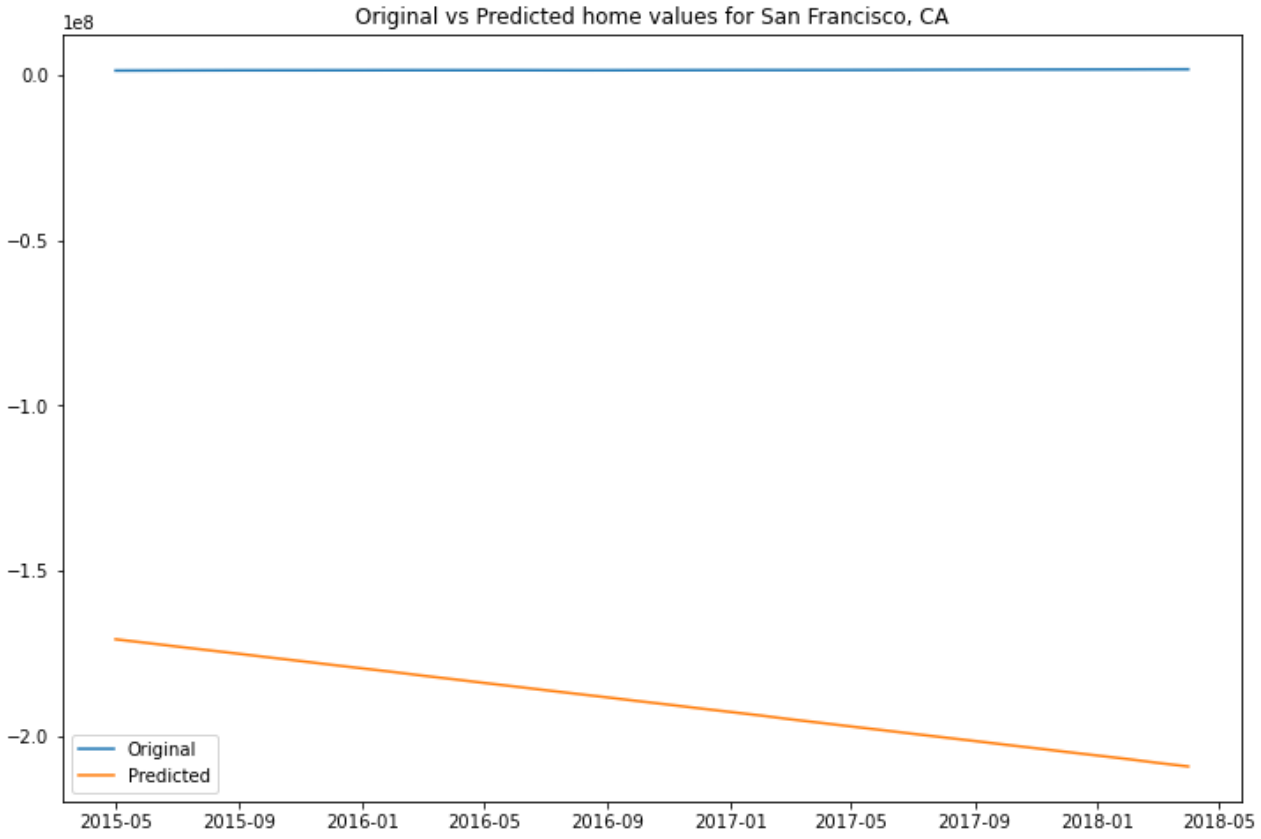
```
---------------------------------------------------------------------------
LinAlgError                               Traceback (most recent call last)
<ipython-input-90-899bb40915ae> in <module>
      4 for city in city_list:
      5     city_model = arima_mod(city)
----> 6     city_model.model(train, test, 4,1,2)
      7     city_model.plot(test)
      8     rmse_list.append(city_model.rmse_)

<ipython-input-66-02957521b05b> in model(self, df_train, df_test, p, d, q)
      7     def model(self, df_train, df_test,p,d,q):
      8         #Fitting our model using ARIMA and instantiating it
----> 9         self.model_fit = ARIMA(df_train[self.city], order = [p,d,q]).fit
()
     10         #Creating our prediction
     11         self.y_hat_test_ = self.model_fit.predict(start=df_test[self.cit
y].index[0],

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/arima/
model.py in fit(self, start_params, transformed, includes_fixed, method, method_
kwargs, gls, gls_kwargs, cov_type, cov_kwds, return_params, low_memory)
    388                 method_kwargs.setdefault('disp', 0)
    389
--> 390                 res = super().fit(
    391                     return_params=return_params, low_memory=low_memory,
    392                     cov_type=cov_type, cov_kwds=cov_kwds, **method_kwarg
s)

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/mlemodel.py in fit(self, start_params, transformed, includes_fixed, cov_typ
e, cov_kwds, method, maxiter, full_output, disp, callback, return_params, optim_
score, optim_complex_step, optim_hessian, flags, low_memory, **kwargs)
    702                 flags['hessian_method'] = optim_hessian
    703             fargs = (flags,)
--> 704             mlefit = super(MLEModel, self).fit(start_params, method=meth
od,
    705                                                fargs=fargs,
    706                                                maxiter=maxiter,

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/base/mode
l.py in fit(self, start_params, method, maxiter, full_output, disp, fargs, callb
ack, retall, skip_hessian, **kwargs)
    561
    562         optimizer = Optimizer()
--> 563         xopt, retvals, optim_settings = optimizer._fit(f, score, start_p
arams,
    564                                                        fargs, kwargs,
    565                                                        hessian=hess,

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/base/optim
izer.py in _fit(self, objective, gradient, start_params, fargs, kwargs, hessian,
method, maxiter, full_output, disp, callback, retall)
    239
    240         func = fit_funcs[method]
--> 241         xopt, retvals = func(objective, gradient, start_params, fargs, k
wargs,
    242                              disp=disp, maxiter=maxiter, callback=callba
ck,
    243                              retall=retall, full_output=full_output,
```

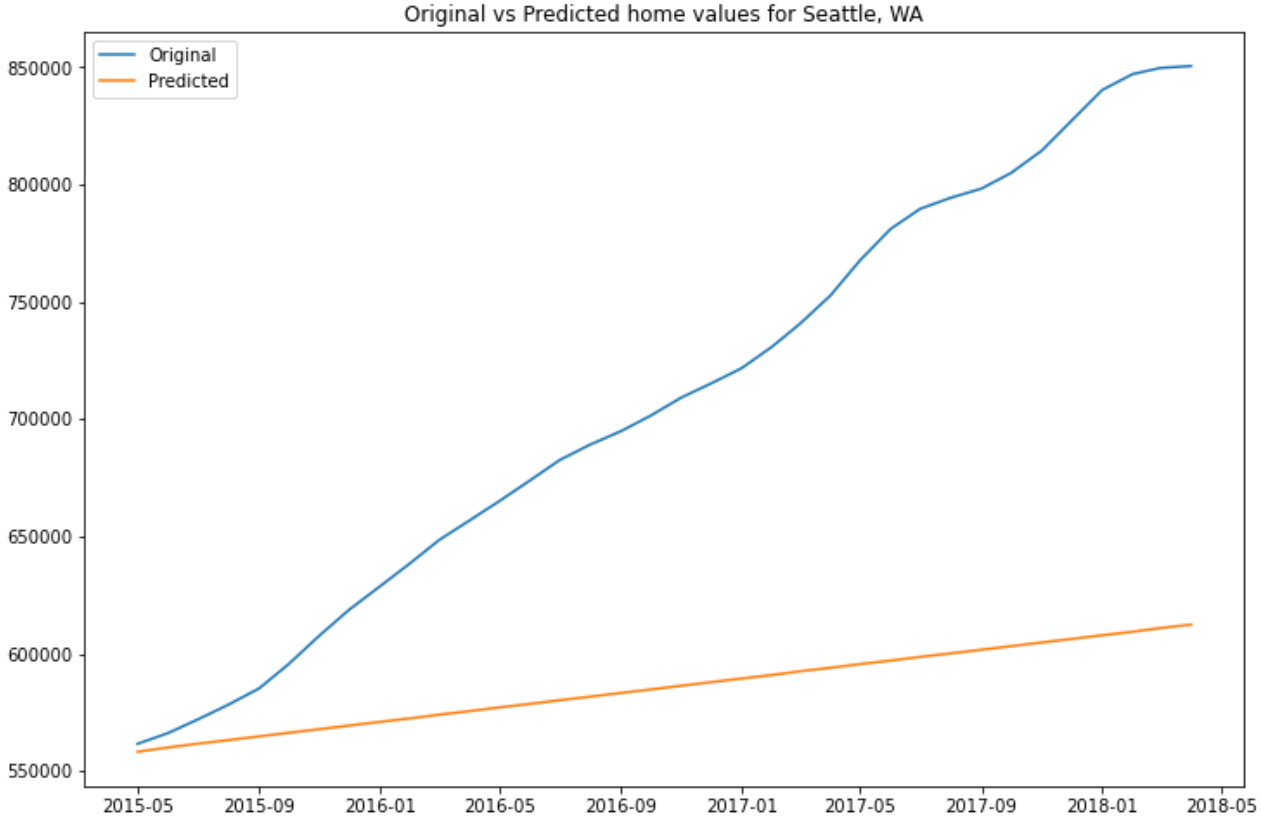
```
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/base/optim
izer.py in _fit_lbfgs(f, score, start_params, fargs, kwargs, disp, maxiter, call
back, retall, full_output, hess)
    649         func = f
    650
--> 651     retvals = optimize.fmin_l_bfgs_b(func, start_params, maxiter=maxite
r,
    652                                      callback=callback, args=fargs,
    653                                      bounds=bounds, disp=disp,

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/scipy/optimize/lbfgsb.
py in fmin_l_bfgs_b(func, x0, fprime, args, approx_grad, bounds, m, factr, pgto
l, epsilon, iprint, maxfun, maxiter, disp, callback, maxls)
    195             'maxls': maxls}
    196
--> 197     res = _minimize_lbfgsb(fun, x0, args=args, jac=jac, bounds=bounds,
    198                            **opts)
    199     d = {'grad': res['jac'],

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/scipy/optimize/lbfgsb.
py in _minimize_lbfgsb(fun, x0, args, jac, bounds, disp, maxcor, ftol, gtol, ep
s, maxfun, maxiter, iprint, callback, maxls, finite_diff_rel_step, **unknown_opt
ions)
    358             # until the completion of the current minimization iteratio
n.
    359             # Overwrite f and g:
--> 360             f, g = func_and_grad(x)
    361         elif task_str.startswith(b'NEW_X'):
    362             # new iteration

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/scipy/optimize/_differ
entiable_functions.py in fun_and_grad(self, x)
    198         if not np.array_equal(x, self.x):
    199             self._update_x_impl(x)
--> 200         self._update_fun()
    201         self._update_grad()
    202         return self.f, self.g

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/scipy/optimize/_differ
entiable_functions.py in _update_fun(self)
    164     def _update_fun(self):
    165         if not self.f_updated:
--> 166             self._update_fun_impl()
    167             self.f_updated = True
    168

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/scipy/optimize/_differ
entiable_functions.py in update_fun()
     71
     72         def update_fun():
---> 73             self.f = fun_wrapped(self.x)
     74
     75         self._update_fun_impl = update_fun

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/scipy/optimize/_differ
entiable_functions.py in fun_wrapped(x)
     68         def fun_wrapped(x):
     69             self.nfev += 1
---> 70             return fun(x, *args)
     71
     72         def update_fun():

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/base/mode
l.py in f(params, *args)
```

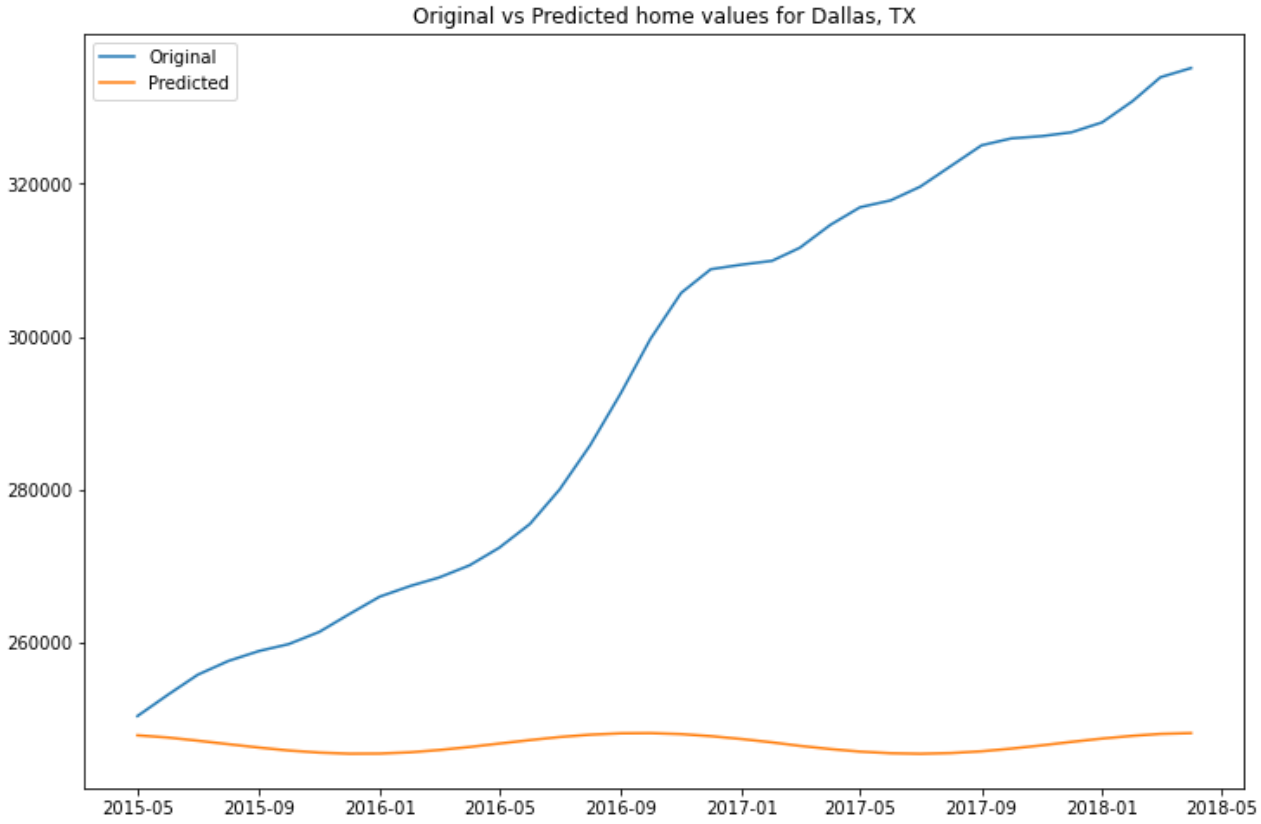
```
    529
    530         def f(params, *args):
--> 531             return -self.loglike(params, *args) / nobs
    532
    533         if method == 'newton':

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/mlemodel.py in loglike(self, params, *args, **kwargs)
    937             kwargs['inversion_method'] = INVERT_UNIVARIATE | SOLVE_LU
    938
--> 939         loglike = self.ssm.loglike(complex_step=complex_step, **kwargs)
    940
    941         # Koopman, Shephard, and Doornik recommend maximizing the averag
e

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/kalman_filter.py in loglike(self, **kwargs)
    981         kwargs.setdefault('conserve_memory',
    982                           MEMORY_CONSERVE ^ MEMORY_NO_LIKELIHOOD)
--> 983         kfilter = self._filter(**kwargs)
    984         loglikelihood_burn = kwargs.get('loglikelihood_burn',
    985                                         self.loglikelihood_burn)

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/kalman_filter.py in _filter(self, filter_method, inversion_method, stabilit
y_method, conserve_memory, filter_timing, tolerance, loglikelihood_burn, complex
_step)
    901
    902         # Initialize the state
--> 903         self._initialize_state(prefix=prefix, complex_step=complex_step)
    904
    905         # Run the filter

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/representation.py in _initialize_state(self, prefix, complex_step)
    981             if not self.initialization.initialized:
    982                 raise RuntimeError('Initialization is incomplete.')
--> 983             self._statespaces[prefix].initialize(self.initialization,
    984                                                  complex_step=complex_st
ep)
    985         else:

statsmodels/tsa/statespace/_representation.pyx in statsmodels.tsa.statespace._re
presentation.dStatespace.initialize()

statsmodels/tsa/statespace/_representation.pyx in statsmodels.tsa.statespace._re
presentation.dStatespace.initialize()

statsmodels/tsa/statespace/_initialization.pyx in statsmodels.tsa.statespace._in
itialization.dInitialization.initialize()

statsmodels/tsa/statespace/_initialization.pyx in statsmodels.tsa.statespace._in
itialization.dInitialization.initialize_stationary_stationary_cov()

statsmodels/tsa/statespace/_tools.pyx in statsmodels.tsa.statespace._tools._dsol
ve_discrete_lyapunov()

LinAlgError: LU decomposition error.
```

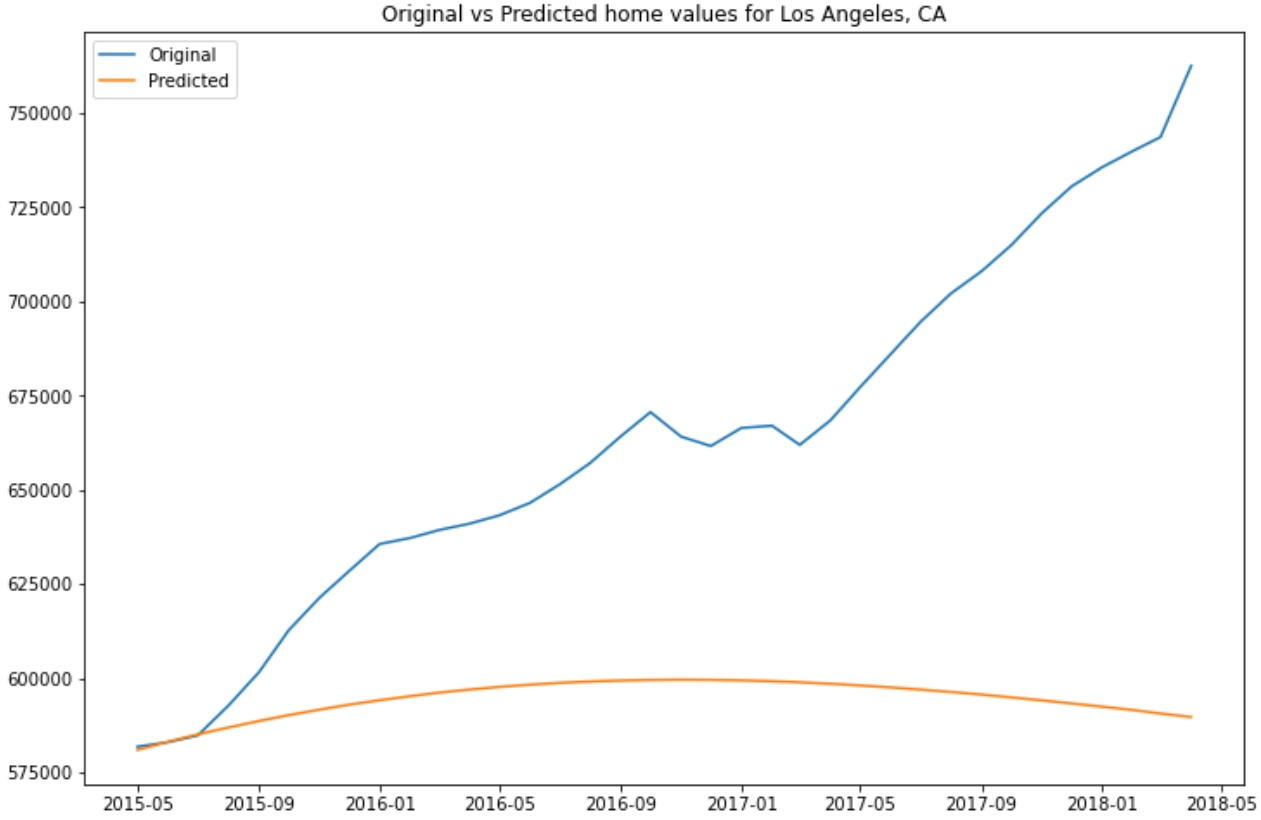

Original vs Predicted home values for Washington, DC
Original
Predicted
780000
760000
740000
720000
700000
680000
660000
640000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for New York, NY
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
550000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for San Francisco, CA
1e8
0.0
−0.5
−1.0
−1.5
−2.0
Original
Predicted
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Seattle, WA
Original
Predicted
850000
800000
750000
700000
650000
600000
550000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05





In [85]:
```
# really rough set of graphs
np.mean(rmse_list)
```

Out[85]:
```
32031328.727358248
```

In [86]:
```
# try and give the moving average some more weight
rmse_list = []
```

```
for city in city_list:
    city_model = arima_mod(city)
    city_model.model(train, test, 4,1,4)
    city_model.plot(test)
    rmse_list.append(city_model.rmse_)
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:         Washington, DC   No. Observations:                  229
Model:                 ARIMA(4, 1, 4)   Log Likelihood               -2840.124
Date:                Fri, 13 May 2022   AIC                           5698.247
Time:                        12:04:12   BIC                           5729.111
Sample:                    04-01-1996   HQIC                          5710.700
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0471      0.011     -4.286      0.000      -0.069      -0.026
ar.L2          1.7325      0.012    149.072      0.000       1.710       1.755
ar.L3          0.1835      0.011     16.996      0.000       0.162       0.205
ar.L4         -0.8688      0.008   -104.969      0.000      -0.885      -0.853
ma.L1          0.2501      0.013     19.641      0.000       0.225       0.275
ma.L2         -1.5485      0.012   -132.201      0.000      -1.571      -1.526
ma.L3         -0.3706      0.012    -32.077      0.000      -0.393      -0.348
ma.L4          0.6700      0.008     84.602      0.000       0.654       0.686
sigma2      4.952e+05   1.62e-08   3.06e+13      0.000    4.95e+05    4.95e+05
===================================================================================
===
Ljung-Box (L1) (Q):                 147.88   Jarque-Bera (JB):                86
0.32
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.02   Skew:                             -
1.44
Prob(H) (two-sided):                  0.94   Kurtosis:                          1
2.07
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 2.32e+
28. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 38428.3794946017
                               SARIMAX Results
==============================================================================
Dep. Variable:           New York, NY   No. Observations:                  229
Model:                 ARIMA(4, 1, 4)   Log Likelihood               -2222.872
Date:                Fri, 13 May 2022   AIC                           4463.745
Time:                        12:04:13   BIC                           4494.609
Sample:                    04-01-1996   HQIC                          4476.197
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.4371      2.574      0.558      0.577      -3.609       6.483
ar.L2         -0.6984      5.081     -0.137      0.891     -10.656       9.259
ar.L3         -0.1956      4.395     -0.045      0.965      -8.809       8.418
ar.L4          0.4013      1.798      0.223      0.823      -3.123       3.926
ma.L1         -1.3787      2.574     -0.536      0.592      -6.424       3.667
```

```
ma.L2          0.6509      4.930      0.132      0.895      -9.011      10.313
ma.L3          0.2062      4.193      0.049      0.961      -8.013       8.425
ma.L4         -0.3806      1.681     -0.226      0.821      -3.676       2.915
sigma2      7.237e+06   1.15e-05   6.28e+11      0.000    7.24e+06    7.24e+06
===================================================================================
===
Ljung-Box (L1) (Q):                  92.98   Jarque-Bera (JB):                 5
7.81
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               2.87   Skew:                            -
0.12
Prob(H) (two-sided):                  0.00   Kurtosis:
5.46
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 6.68e+
27. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 93770.72873801329
                               SARIMAX Results
===================================================================================
Dep. Variable:      San Francisco, CA   No. Observations:                   229
Model:                 ARIMA(4, 1, 4)   Log Likelihood               -2722.050
Date:                Fri, 13 May 2022   AIC                           5462.100
Time:                        12:04:15   BIC                           5492.964
Sample:                    04-01-1996   HQIC                          5474.553
                         - 04-01-2015
Covariance Type:                  opg
===================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
ar.L1         -0.1188      0.018     -6.624      0.000      -0.154      -0.084
ar.L2          1.7845      0.018     97.118      0.000       1.748       1.820
ar.L3          0.1775      0.020      8.666      0.000       0.137       0.218
ar.L4         -0.8432      0.019    -44.299      0.000      -0.880      -0.806
ma.L1          0.1920      0.019     10.309      0.000       0.156       0.229
ma.L2         -1.7126      0.018    -92.636      0.000      -1.749      -1.676
ma.L3         -0.2478      0.021    -11.813      0.000      -0.289      -0.207
ma.L4          0.7696      0.019     40.320      0.000       0.732       0.807
sigma2      7.028e+06   1.76e-09      4e+15      0.000    7.03e+06    7.03e+06
===================================================================================
===
Ljung-Box (L1) (Q):                 165.33   Jarque-Bera (JB):                 1
0.62
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               2.25   Skew:                            -
0.36
Prob(H) (two-sided):                  0.00   Kurtosis:
3.77
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 2.03e+
31. Standard errors may be unstable.
-----------------------
```

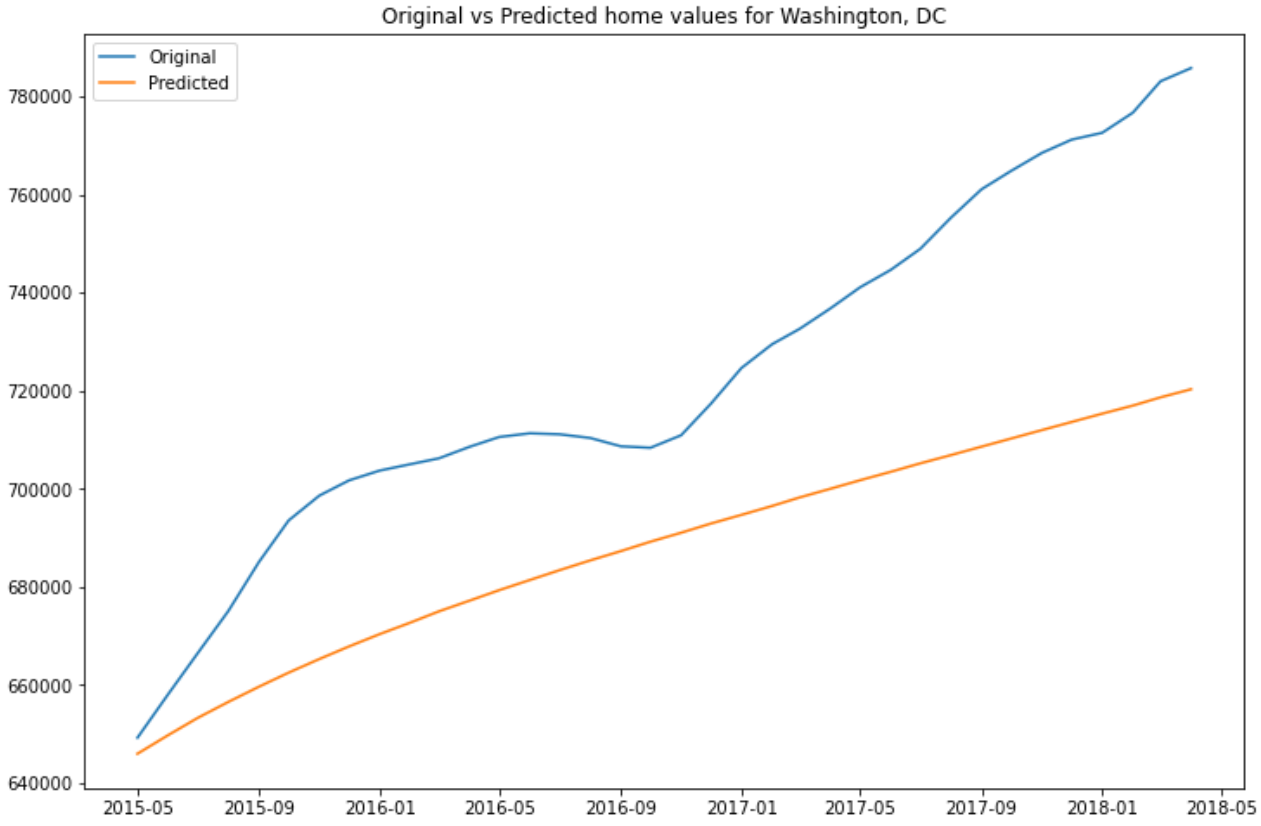
```
-----------------------
RMSE: 120855.59354723897
                               SARIMAX Results
==============================================================================
Dep. Variable:            Seattle, WA   No. Observations:                  229
Model:                 ARIMA(4, 1, 4)   Log Likelihood               -2676.124
Date:                Fri, 13 May 2022   AIC                           5370.248
Time:                        12:04:16   BIC                           5401.112
Sample:                    04-01-1996   HQIC                          5382.701
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7288      2.040      0.357      0.721      -3.269       4.727
ar.L2          0.8601      0.825      1.042      0.297      -0.757       2.477
ar.L3         -0.7453      1.959     -0.380      0.704      -4.585       3.095
ar.L4          0.1156      0.878      0.132      0.895      -1.606       1.837
ma.L1         -0.6048      2.040     -0.297      0.767      -4.603       3.393
ma.L2         -0.8576      0.574     -1.495      0.135      -1.982       0.267
ma.L3          0.6503      1.871      0.348      0.728      -3.016       4.317
ma.L4         -0.0917      0.720     -0.127      0.899      -1.504       1.320
sigma2      6.495e+05      0.000   3.05e+09      0.000     6.5e+05     6.5e+05
===================================================================================
===
Ljung-Box (L1) (Q):                 160.71   Jarque-Bera (JB):                 14
8.88
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.90   Skew:                             -
1.20
Prob(H) (two-sided):                  0.01   Kurtosis:
6.15
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 9.01e+
24. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 146929.67137701577
                               SARIMAX Results
==============================================================================
Dep. Variable:              Dallas, TX   No. Observations:                  229
Model:                 ARIMA(4, 1, 4)   Log Likelihood               -1945.753
Date:                Fri, 13 May 2022   AIC                           3909.507
Time:                        12:04:16   BIC                           3940.371
Sample:                    04-01-1996   HQIC                          3921.960
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.1878      3.220      0.369      0.712      -5.123       7.499
ar.L2          0.0765      6.826      0.011      0.991     -13.302      13.455
ar.L3         -0.6533      5.837     -0.112      0.911     -12.094      10.787
ar.L4          0.1394      1.992      0.070      0.944      -3.766       4.044
ma.L1         -1.1631      3.220     -0.361      0.718      -7.475       5.149
ma.L2         -0.0894      6.754     -0.013      0.989     -13.326      13.147
ma.L3          0.6416      5.728      0.112      0.911     -10.584      11.868
ma.L4         -0.1254      1.939     -0.065      0.948      -3.926       3.675
sigma2      1.257e+06   2.05e+04     61.394      0.000    1.22e+06     1.3e+06
```

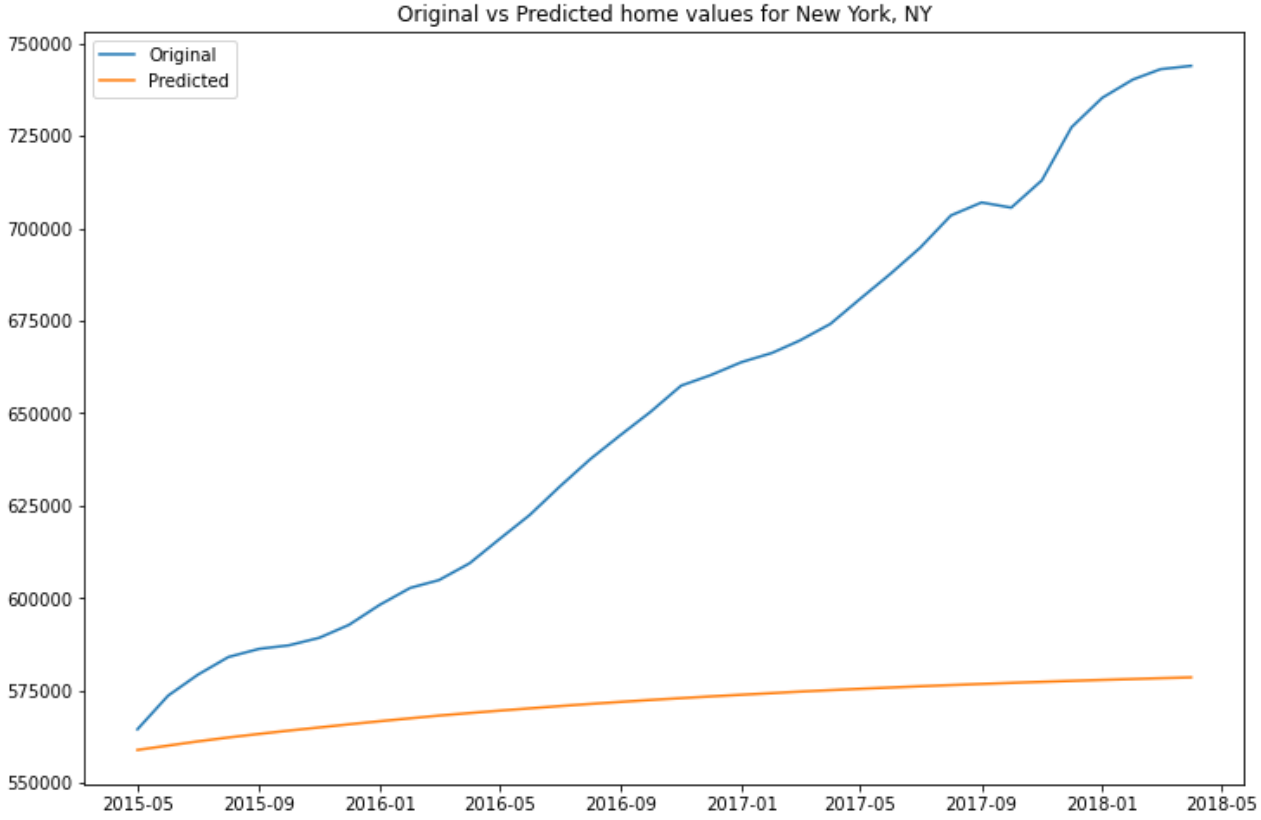
```
==========================================================================
===
Ljung-Box (L1) (Q):                  27.79   Jarque-Bera (JB):             4519
6.29
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              11.32   Skew:                           -
5.89
Prob(H) (two-sided):                  0.00   Kurtosis:                        7
0.96
==========================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
-----------------------
-----------------------
RMSE: 54499.66683677033
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:966: UserWarning:

Non-stationary starting autoregressive parameters found. Using zeros as starting
parameters.

/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:978: UserWarning:

Non-invertible starting MA parameters found. Using zeros as starting parameters.

                               SARIMAX Results
==========================================================================
Dep. Variable:        Los Angeles, CA   No. Observations:                  229
Model:                ARIMA(4, 1, 4)   Log Likelihood               -2322.046
Date:               Fri, 13 May 2022   AIC                         4662.093
Time:                       12:04:17   BIC                         4692.957
Sample:                   04-01-1996   HQIC                        4674.545
                        - 04-01-2015
Covariance Type:                 opg
==========================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------
ar.L1          0.7409      0.168      4.418      0.000       0.412       1.070
ar.L2          1.0056      0.248      4.049      0.000       0.519       1.492
ar.L3         -0.6974      0.172     -4.053      0.000      -1.035      -0.360
ar.L4         -0.0774      0.119     -0.649      0.516      -0.311       0.156
ma.L1         -0.6561      0.168     -3.903      0.000      -0.986      -0.327
ma.L2         -1.0307      0.237     -4.352      0.000      -1.495      -0.566
ma.L3          0.6283      0.163      3.855      0.000       0.309       0.948
ma.L4          0.1156      0.112      1.036      0.300      -0.103       0.334
sigma2      4.995e+06   2.24e-07   2.23e+13      0.000       5e+06       5e+06
==========================================================================
===
Ljung-Box (L1) (Q):                 137.91   Jarque-Bera (JB):               3
7.27
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.28   Skew:                           -
0.44
Prob(H) (two-sided):                  0.29   Kurtosis:
4.78
==========================================================================
===

Warnings:
```

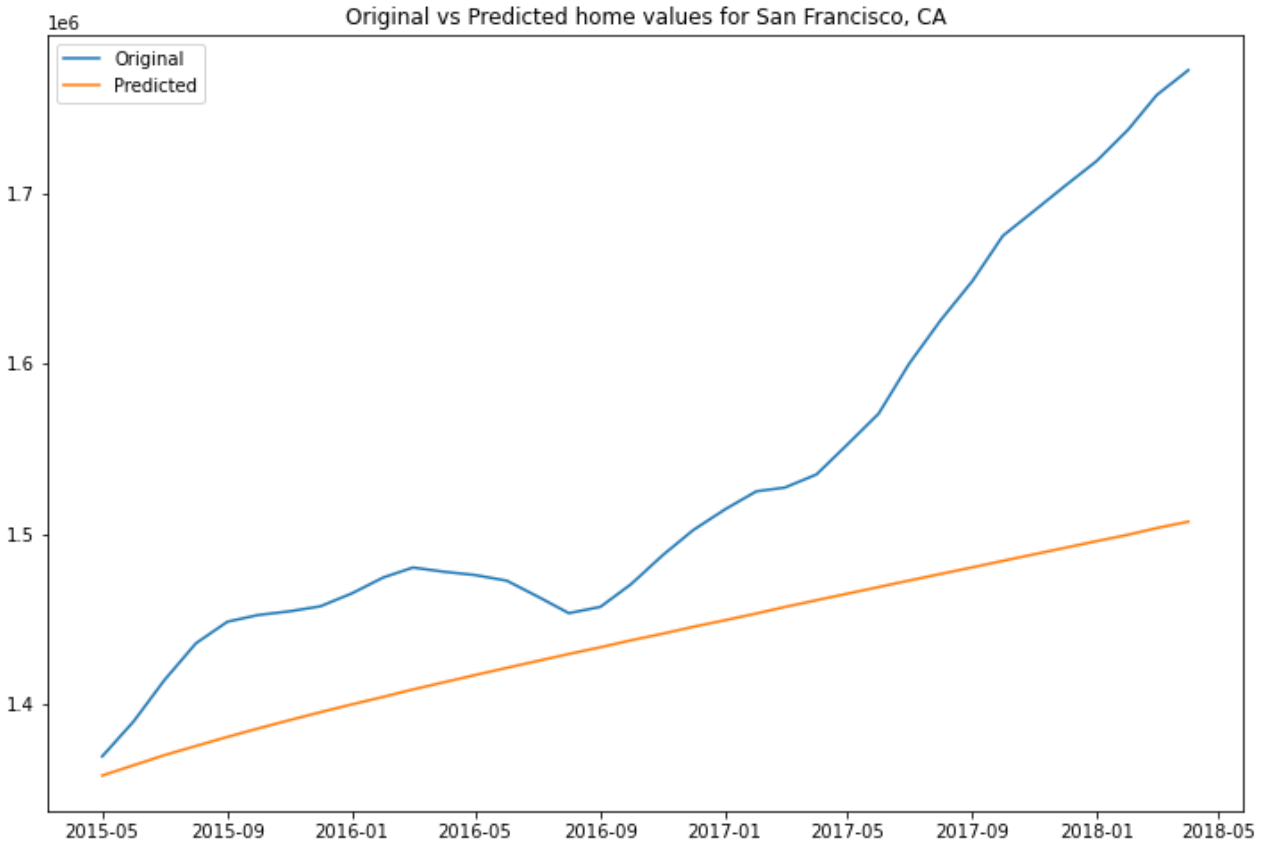
```
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 8.36e+
28. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 80131.90124457421
                               SARIMAX Results
==============================================================================
Dep. Variable:           San Jose, CA   No. Observations:                  229
Model:                 ARIMA(4, 1, 4)   Log Likelihood               -2630.601
Date:                Fri, 13 May 2022   AIC                           5279.201
Time:                        12:04:19   BIC                           5310.065
Sample:                    04-01-1996   HQIC                          5291.654
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0620      0.048     -1.285      0.199      -0.156       0.033
ar.L2          1.8352      0.047     38.952      0.000       1.743       1.928
ar.L3          0.1136      0.045      2.515      0.012       0.025       0.202
ar.L4         -0.8870      0.043    -20.813      0.000      -0.970      -0.803
ma.L1          0.1353      0.049      2.790      0.005       0.040       0.230
ma.L2         -1.7650      0.044    -39.931      0.000      -1.852      -1.678
ma.L3         -0.1845      0.046     -4.028      0.000      -0.274      -0.095
ma.L4          0.8149      0.040     20.350      0.000       0.736       0.893
sigma2      2.954e+06   9.25e-09   3.19e+14      0.000    2.95e+06    2.95e+06
==============================================================================
===
Ljung-Box (L1) (Q):                 172.65   Jarque-Bera (JB):                 1
8.41
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.29   Skew:                            -
0.39
Prob(H) (two-sided):                  0.28   Kurtosis:
4.15
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 6.23e+
31. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 119125.2511290605
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:966: UserWarning:

Non-stationary starting autoregressive parameters found. Using zeros as starting
parameters.

                               SARIMAX Results
==============================================================================
Dep. Variable:            Chicago, IL   No. Observations:                  229
Model:                 ARIMA(4, 1, 4)   Log Likelihood               -2319.835
Date:                Fri, 13 May 2022   AIC                           4657.670
Time:                        12:04:20   BIC                           4688.534
Sample:                    04-01-1996   HQIC                          4670.122
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
```

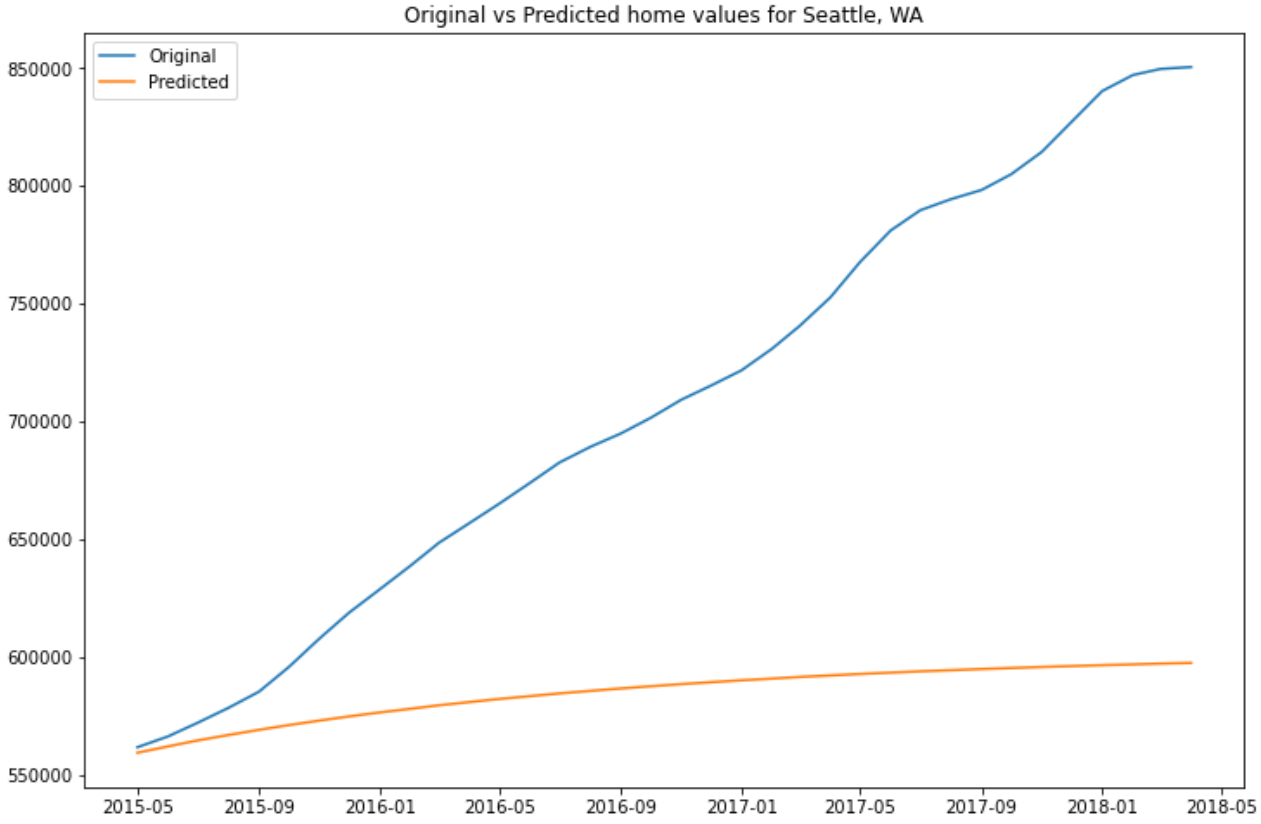
```
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.2791      1.590     -0.176      0.861      -3.395       2.837
ar.L2          1.1119      0.635      1.752      0.080      -0.132       2.356
ar.L3          0.3097      1.341      0.231      0.817      -2.318       2.937
ar.L4         -0.2619      0.497     -0.527      0.598      -1.237       0.713
ma.L1          0.3920      1.590      0.247      0.805      -2.724       3.508
ma.L2         -1.0016      0.460     -2.179      0.029      -1.903      -0.100
ma.L3         -0.3555      1.285     -0.277      0.782      -2.874       2.163
ma.L4          0.2096      0.388      0.540      0.589      -0.551       0.971
sigma2      7.174e+05    1.1e+04     65.302      0.000    6.96e+05    7.39e+05
===================================================================================
===
Ljung-Box (L1) (Q):                 157.18   Jarque-Bera (JB):                21
2.15
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.32   Skew:                             -
0.43
Prob(H) (two-sided):                  0.22   Kurtosis:
7.65
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 6.7e+1
4. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 23551.49464533649
                               SARIMAX Results
==============================================================================
Dep. Variable:          Baltimore, MD   No. Observations:                  229
Model:                 ARIMA(4, 1, 4)   Log Likelihood               -1897.493
Date:                Fri, 13 May 2022   AIC                           3812.985
Time:                        12:04:21   BIC                           3843.850
Sample:                    04-01-1996   HQIC                          3825.438
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          2.0592      0.206      9.992      0.000       1.655       2.463
ar.L2         -2.0328      0.407     -4.996      0.000      -2.830      -1.235
ar.L3          0.9441      0.366      2.583      0.010       0.228       1.661
ar.L4         -0.0238      0.150     -0.159      0.874      -0.318       0.270
ma.L1         -1.7156      0.208     -8.263      0.000      -2.123      -1.309
ma.L2          1.4772      0.350      4.220      0.000       0.791       2.163
ma.L3         -0.4452      0.291     -1.531      0.126      -1.015       0.125
ma.L4         -0.1390      0.107     -1.303      0.193      -0.348       0.070
sigma2      2.749e+05   1.24e+04     22.184      0.000    2.51e+05    2.99e+05
===================================================================================
===
Ljung-Box (L1) (Q):                  73.97   Jarque-Bera (JB):               123
4.66
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               1.77   Skew:                             -
1.69
Prob(H) (two-sided):                  0.01   Kurtosis:                          1
3.89
===================================================================================
===
```

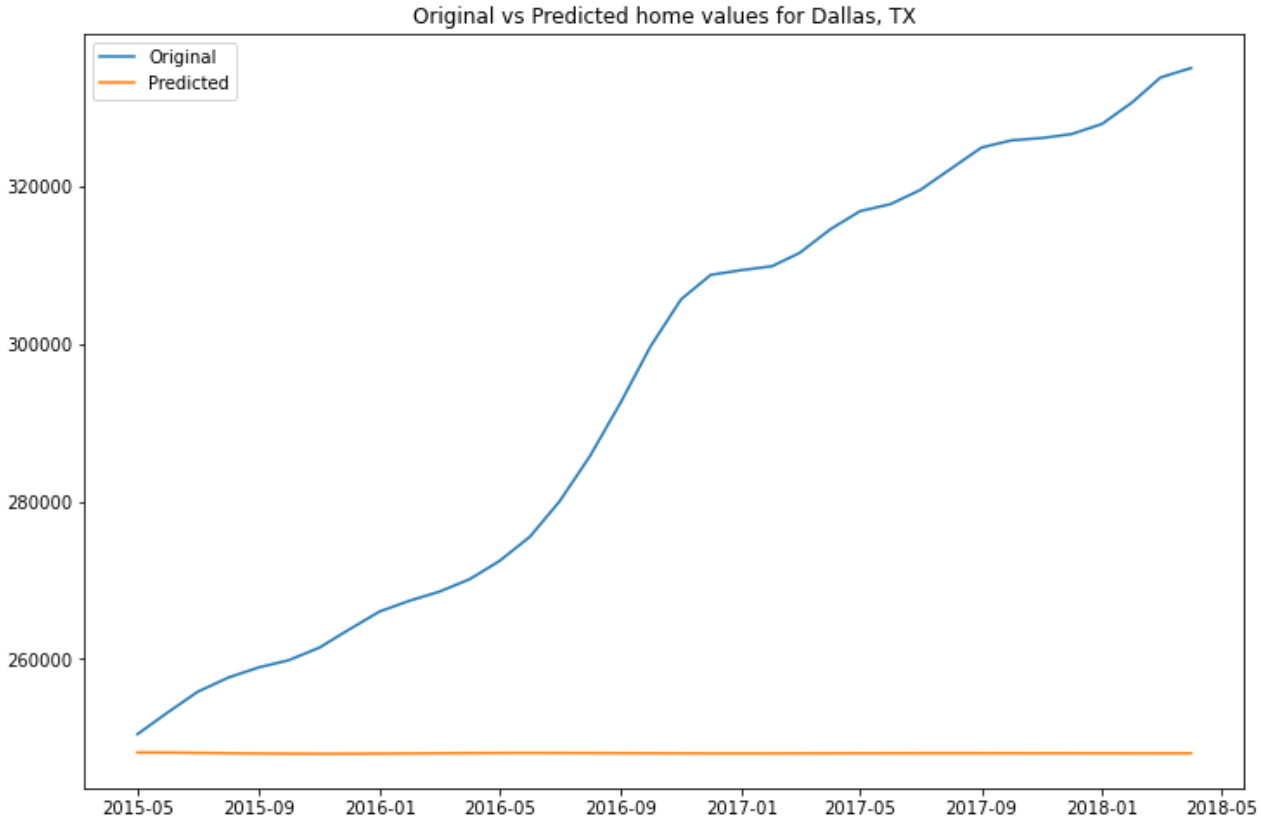
```
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 7.82e+
14. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 9159.344548415105
/opt/anaconda3/envs/learn-env/lib/python3.8/site-packages/statsmodels/tsa/states
pace/sarimax.py:966: UserWarning:

Non-stationary starting autoregressive parameters found. Using zeros as starting
parameters.

                               SARIMAX Results
==============================================================================
Dep. Variable:              Boston, MA   No. Observations:                  229
Model:                 ARIMA(4, 1, 4)   Log Likelihood               -2105.997
Date:                Fri, 13 May 2022   AIC                           4229.995
Time:                        12:04:23   BIC                           4260.859
Sample:                    04-01-1996   HQIC                          4242.448
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.4132      0.330      4.282      0.000       0.766       2.060
ar.L2          0.5267      0.610      0.863      0.388      -0.669       1.723
ar.L3         -1.4139      0.566     -2.496      0.013      -2.524      -0.304
ar.L4          0.4712      0.242      1.951      0.051      -0.002       0.945
ma.L1         -1.3381      0.333     -4.016      0.000      -1.991      -0.685
ma.L2         -0.5820      0.602     -0.967      0.334      -1.762       0.598
ma.L3          1.3404      0.548      2.447      0.014       0.267       2.414
ma.L4         -0.4141      0.227     -1.823      0.068      -0.859       0.031
sigma2      2.325e+06   5.88e-06   3.95e+11      0.000    2.32e+06    2.32e+06
==============================================================================
===
Ljung-Box (L1) (Q):                 101.43   Jarque-Bera (JB):               20
4.15
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               2.19   Skew:
0.43
Prob(H) (two-sided):                  0.00   Kurtosis:
7.55
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
[2] Covariance matrix is singular or near-singular, with condition number 7.03e+
26. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 73661.07813482023
```

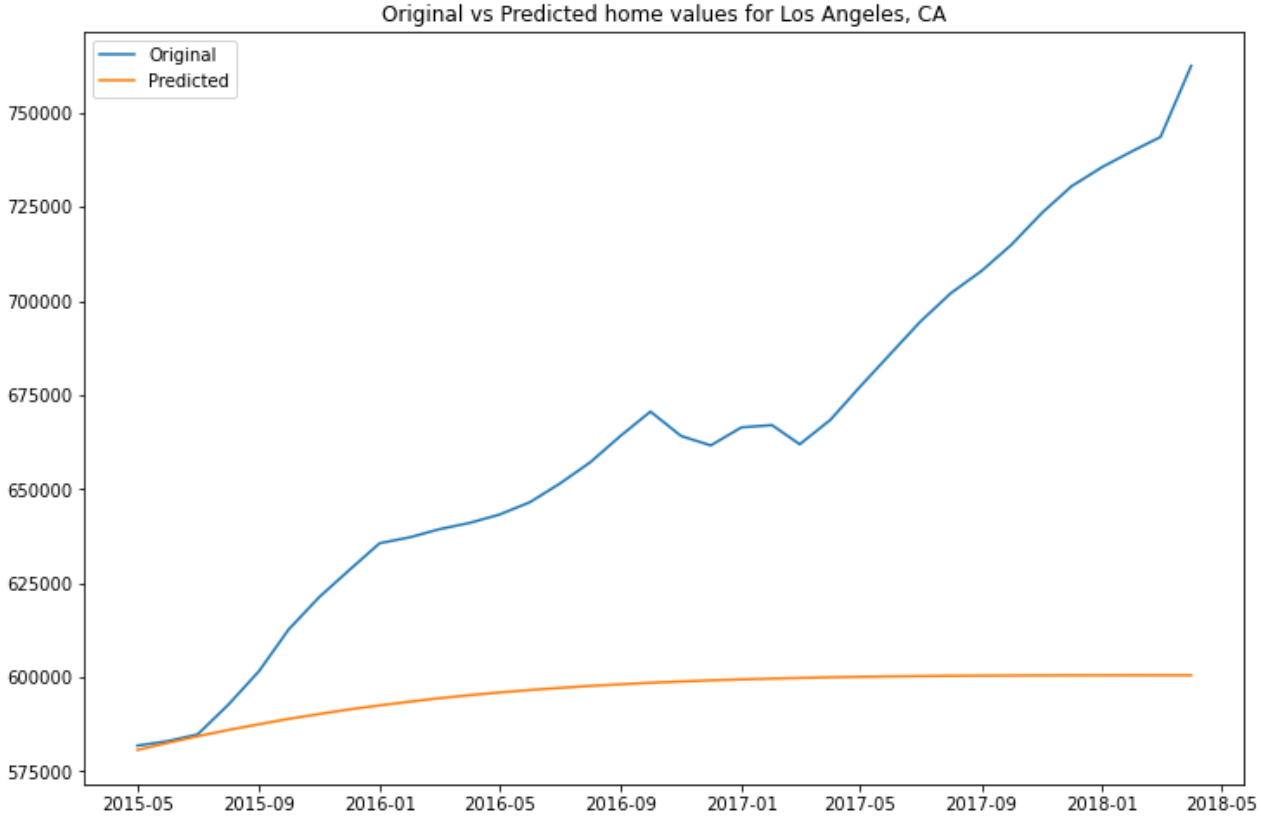

Original vs Predicted home values for Washington, DC
Original
Predicted
780000
760000
740000
720000
700000
680000
660000
640000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

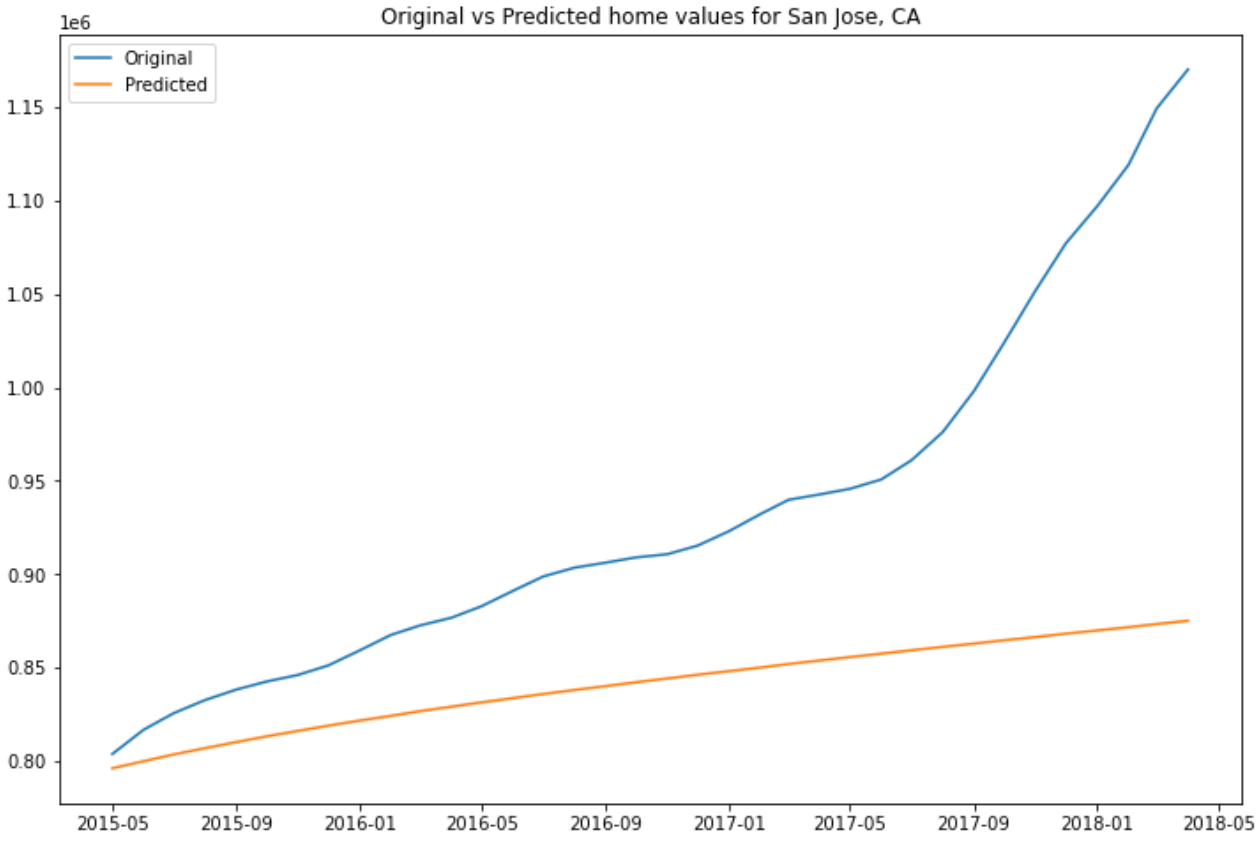

Original vs Predicted home values for New York, NY
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
550000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

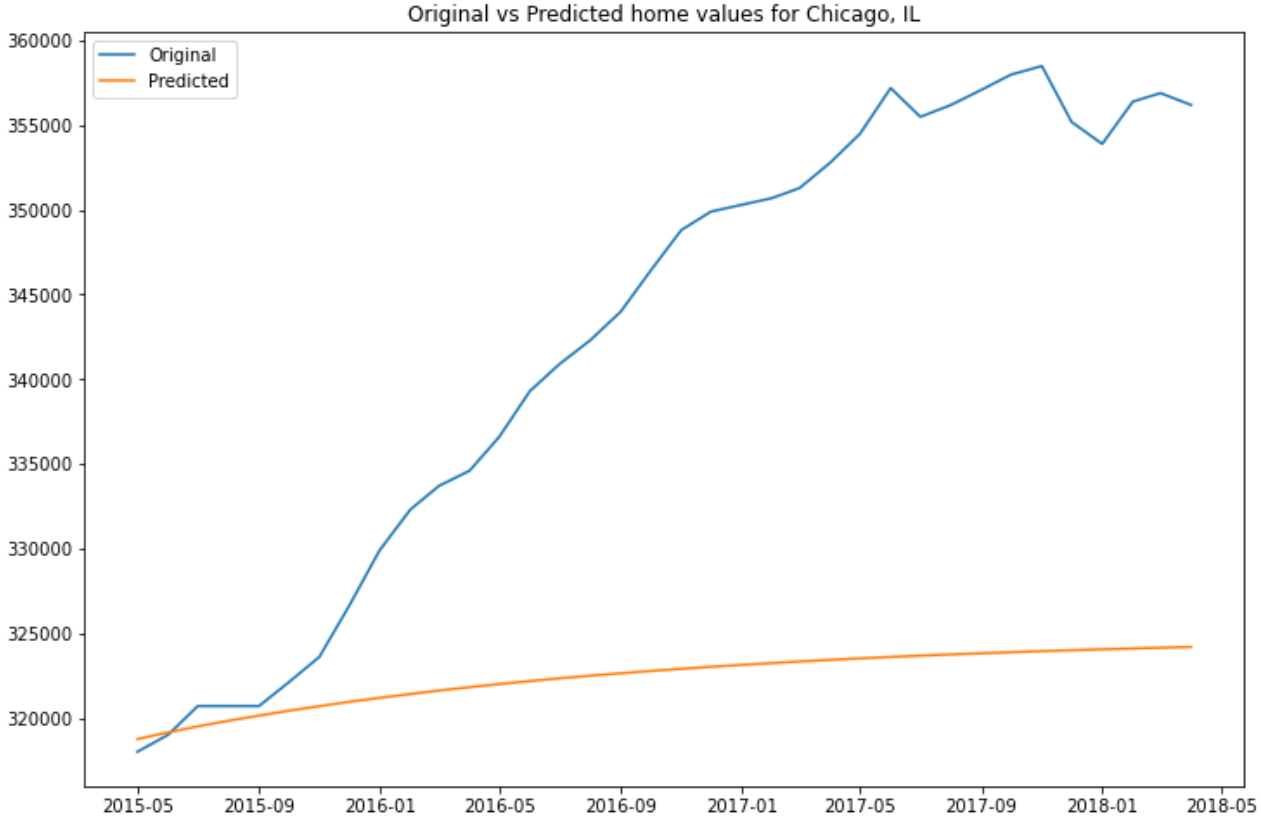

Original vs Predicted home values for San Francisco, CA
1e6
Original
Predicted
1.7
1.6
1.5
1.4
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

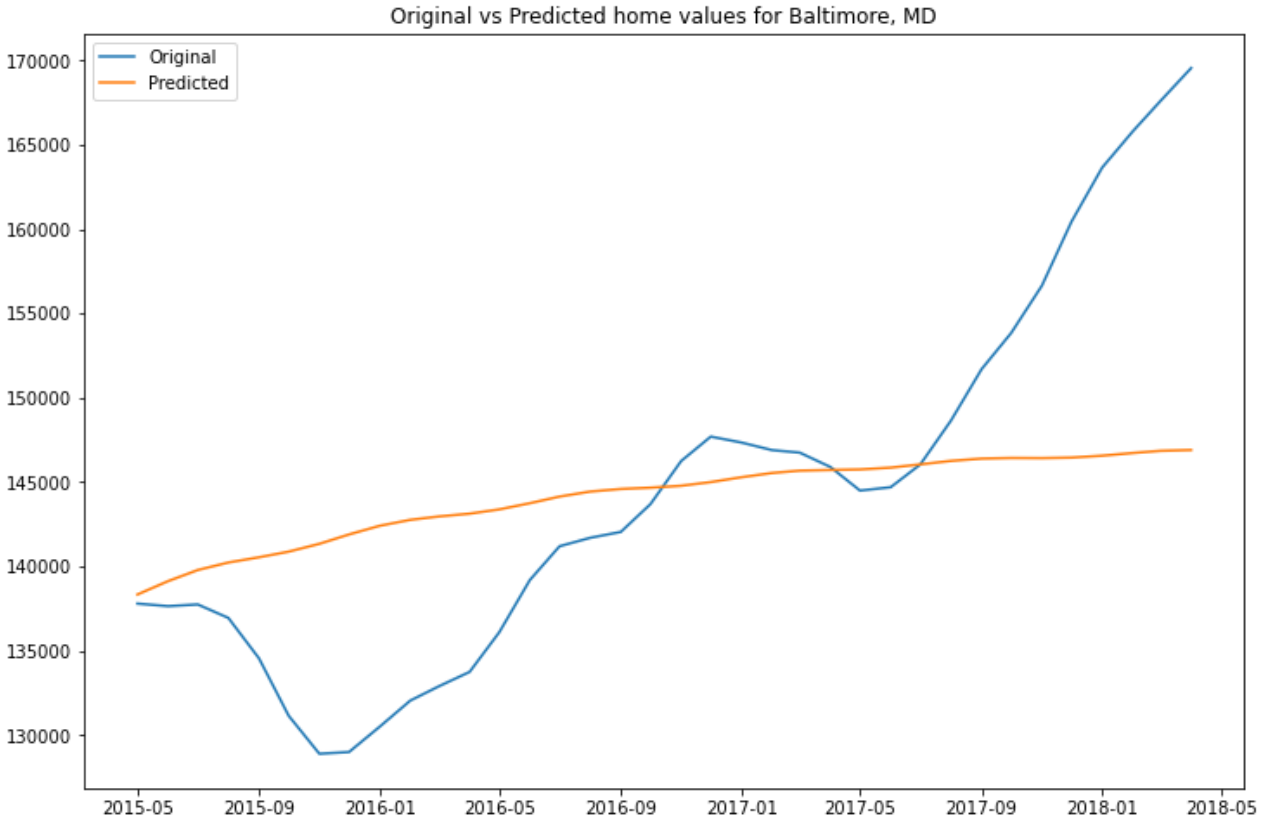

Original vs Predicted home values for Seattle, WA
Original
Predicted
850000
800000
750000
700000
650000
600000
550000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05

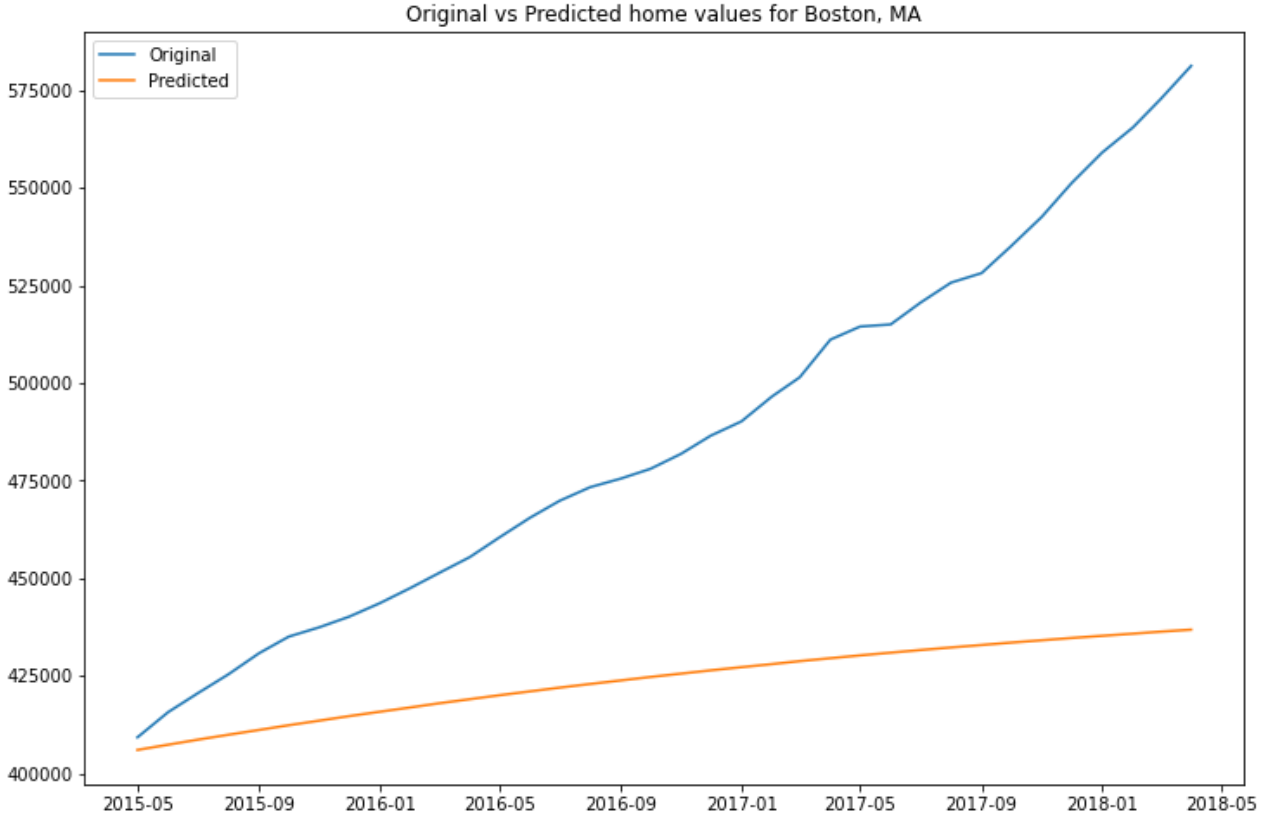

Original vs Predicted home values for Dallas, TX
Original
Predicted
320000
300000
280000
260000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Los Angeles, CA
Original
Predicted
750000
725000
700000
675000
650000
625000
600000
575000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for San Jose, CA
1e6
Original
Predicted
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05


Original vs Predicted home values for Chicago, IL
Original
Predicted
360000
355000
350000
345000
340000
335000
330000
325000
320000
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05





In [91]:
```
# still rough, don't know if much improvement can be made from 1,2,3 for p,d,q
np.mean(rmse_list)
```

Out[91]:
```
32031328.727358248
```

Visualization for Model Performance

In [92]:
```
color_list=['blue', 'orange', 'green', 'red', 'purple', 'brown', 'pink', 'gray',
fig, ax = plt.subplots(figsize = (20,15))
for city in city_list:
    city_model = arima_mod(city)
    city_model.model(train, test, 1,2,3)
    ax.plot(test[city], color=color_list[city_list.index(city)])
    ax.plot(city_model.y_hat_test_, linestyle='--', color=color_list[city_list.i

leg = plt.legend(['Original', 'Predicted'], loc=2)
ax.add_artist(leg)

patch_list=[]
for color,city in zip(color_list,city_list):
    patch = mpatches.Patch(color=color, label=city)
    patch_list.append(patch)
plt.legend(handles=patch_list, loc=1)



ax.set_title('Original vs Predicted home values')

fig.savefig('figures/Original vs Predicted.jpeg', dpi=500)
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:         Washington, DC   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -1857.847
Date:                Fri, 13 May 2022   AIC                           3725.693
Time:                        12:06:34   BIC                           3742.818
Sample:                    04-01-1996   HQIC                          3732.603
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9714      0.023     41.616      0.000       0.926       1.017
ma.L1         -0.9290      0.022    -41.750      0.000      -0.973      -0.885
ma.L2         -0.0345      0.017     -2.056      0.040      -0.067      -0.002
ma.L3         -0.0238      0.024     -1.006      0.314      -0.070       0.023
sigma2      7.277e+05   4.39e+04     16.594      0.000    6.42e+05    8.14e+05
==============================================================================
===
Ljung-Box (L1) (Q):                   49.02   Jarque-Bera (JB):                10
3.80
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):               15.74   Skew:
0.25
Prob(H) (two-sided):                  0.00   Kurtosis:
6.28
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 36748.34494860473
                               SARIMAX Results
==============================================================================
Dep. Variable:           New York, NY   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -2126.309
```

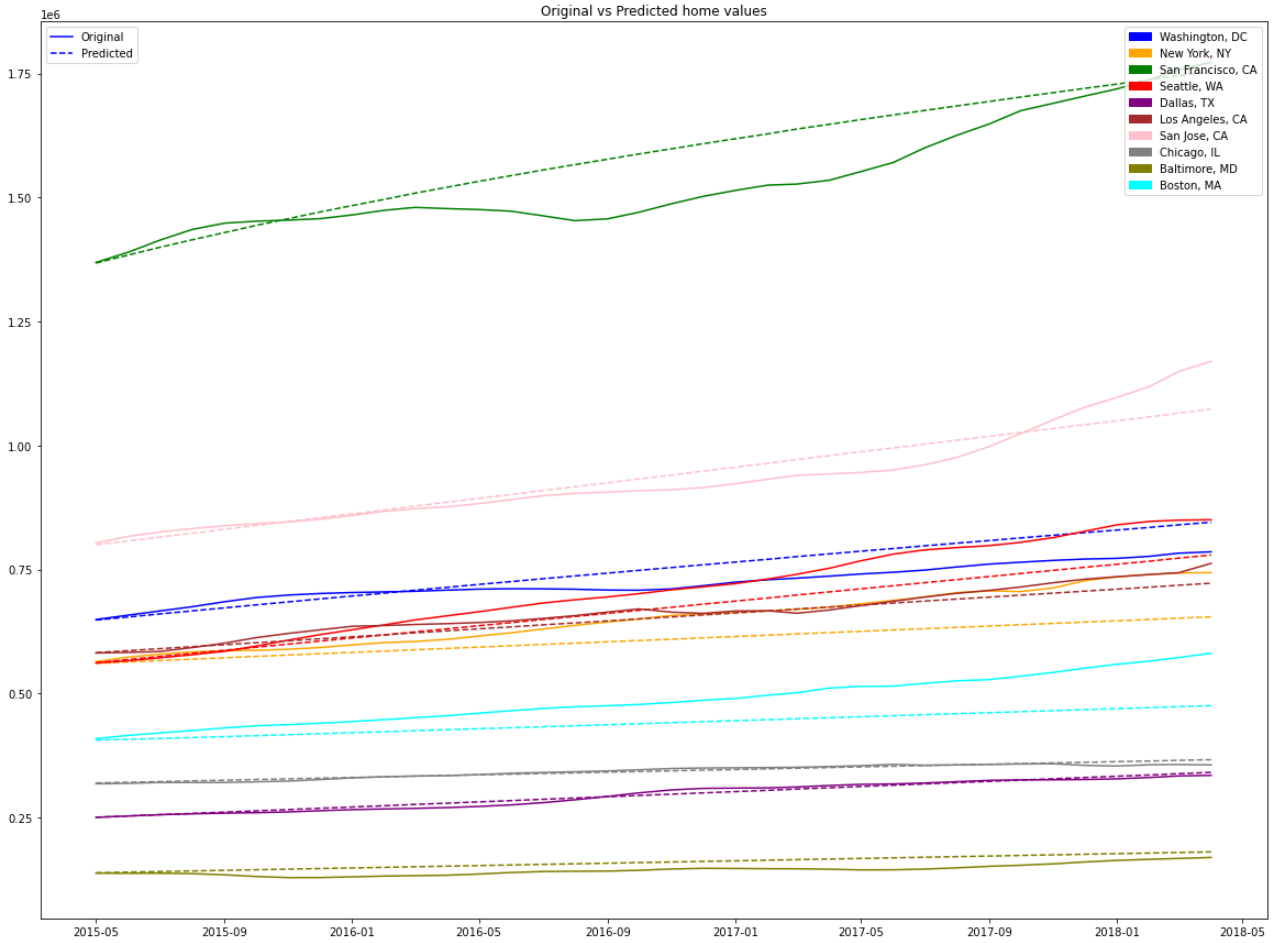
```
Date:                Fri, 13 May 2022   AIC                           4262.617
Time:                        12:06:35   BIC                           4279.742
Sample:                  04-01-1996   HQIC                          4269.528
                       - 04-01-2015
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7494      0.076      9.887      0.000       0.601       0.898
ma.L1         -0.8831      0.093     -9.498      0.000      -1.065      -0.701
ma.L2         -0.0310      0.059     -0.530      0.596      -0.146       0.084
ma.L3         -0.0220      0.027     -0.806      0.420      -0.075       0.031
sigma2      7.341e+06    3.9e+05     18.834      0.000    6.58e+06     8.1e+06
===================================================================================
===
Ljung-Box (L1) (Q):                   1.55   Jarque-Bera (JB):                27
8.23
Prob(Q):                              0.21   Prob(JB):
0.00
Heteroskedasticity (H):              14.04   Skew:                             -
0.21
Prob(H) (two-sided):                  0.00   Kurtosis:
8.41
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
-----------------------
-----------------------
RMSE: 50867.36839003728
                               SARIMAX Results
==============================================================================
Dep. Variable:      San Francisco, CA   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -2142.881
Date:                Fri, 13 May 2022   AIC                           4295.762
Time:                        12:06:35   BIC                           4312.887
Sample:                  04-01-1996   HQIC                          4302.672
                       - 04-01-2015
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9388      0.025     38.299      0.000       0.891       0.987
ma.L1         -0.9759      0.045    -21.771      0.000      -1.064      -0.888
ma.L2         -0.0160      0.011     -1.483      0.138      -0.037       0.005
ma.L3         -0.0070      0.025     -0.280      0.780      -0.056       0.042
sigma2      7.765e+06   2.85e-09   2.73e+15      0.000    7.77e+06    7.77e+06
===================================================================================
===
Ljung-Box (L1) (Q):                  21.12   Jarque-Bera (JB):                9
1.84
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              18.69   Skew:
0.40
Prob(H) (two-sided):                  0.00   Kurtosis:
6.01
===================================================================================
===

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

```
[2] Covariance matrix is singular or near-singular, with condition number 1.27e+
30. Standard errors may be unstable.
-----------------------
-----------------------
RMSE: 67963.99107969219
                               SARIMAX Results
==============================================================================
Dep. Variable:            Seattle, WA   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -1878.103
Date:                Fri, 13 May 2022   AIC                           3766.205
Time:                        12:06:36   BIC                           3783.330
Sample:                    04-01-1996   HQIC                          3773.115
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0221      2.478     -0.009      0.993      -4.879       4.835
ma.L1          0.0281      2.484      0.011      0.991      -4.840       4.896
ma.L2         -0.0146      0.036     -0.411      0.681      -0.084       0.055
ma.L3         -0.0168      0.025     -0.665      0.506      -0.066       0.033
sigma2      8.889e+05   6.53e+04     13.610      0.000    7.61e+05    1.02e+06
==============================================================================
===
Ljung-Box (L1) (Q):                  24.43   Jarque-Bera (JB):                 2
5.50
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              26.23   Skew:
0.28
Prob(H) (two-sided):                  0.00   Kurtosis:
4.54
==============================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
-----------------------
-----------------------
RMSE: 45483.41561208202
                               SARIMAX Results
==============================================================================
Dep. Variable:              Dallas, TX   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -1955.950
Date:                Fri, 13 May 2022   AIC                           3921.900
Time:                        12:06:36   BIC                           3939.025
Sample:                    04-01-1996   HQIC                          3928.810
                         - 04-01-2015
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7822      5.496     -0.142      0.887     -11.553       9.989
ma.L1          0.7501      5.498      0.136      0.891     -10.026      11.526
ma.L2         -0.0381      0.171     -0.223      0.824      -0.373       0.297
ma.L3         -0.0118      0.064     -0.184      0.854      -0.138       0.115
sigma2       1.34e+06   1.82e+04     73.572      0.000     1.3e+06    1.38e+06
==============================================================================
===
Ljung-Box (L1) (Q):                  38.57   Jarque-Bera (JB):              7643
2.64
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              45.20   Skew:                             -
```

```
1.52
Prob(H) (two-sided):                  0.00   Kurtosis:                         9
2.84
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 5165.853590925947
                               SARIMAX Results
===================================================================================
Dep. Variable:        Los Angeles, CA   No. Observations:                   229
Model:                 ARIMA(1, 2, 3)   Log Likelihood                -2075.658
Date:                Fri, 13 May 2022   AIC                             4161.315
Time:                        12:06:36   BIC                             4178.440
Sample:                    04-01-1996   HQIC                            4168.225
                         - 04-01-2015
Covariance Type:                  opg
===================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
ar.L1          0.7735      0.205      3.774      0.000       0.372       1.175
ma.L1         -0.8365      0.209     -3.997      0.000      -1.247      -0.426
ma.L2         -0.0073      0.029     -0.253      0.800      -0.064       0.049
ma.L3         -0.0112      0.026     -0.422      0.673      -0.063       0.041
sigma2      5.032e+06   1.98e+05     25.411      0.000    4.64e+06    5.42e+06
===================================================================================
===
Ljung-Box (L1) (Q):                   8.62   Jarque-Bera (JB):              1543
5.05
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              26.61   Skew:                             -
3.27
Prob(H) (two-sided):                  0.00   Kurtosis:                         4
2.86
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 15666.858520860827
                               SARIMAX Results
===================================================================================
Dep. Variable:           San Jose, CA   No. Observations:                   229
Model:                 ARIMA(1, 2, 3)   Log Likelihood                -2028.183
Date:                Fri, 13 May 2022   AIC                             4066.366
Time:                        12:06:37   BIC                             4083.490
Sample:                    04-01-1996   HQIC                            4073.276
                         - 04-01-2015
Covariance Type:                  opg
===================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
ar.L1         -0.1906      8.385     -0.023      0.982     -16.624      16.243
ma.L1          0.1809      8.396      0.022      0.983     -16.276      16.638
ma.L2         -0.0174      0.069     -0.253      0.800      -0.152       0.118
ma.L3         -0.0074      0.112     -0.066      0.948      -0.227       0.213
sigma2      3.072e+06   1.37e+05     22.457      0.000     2.8e+06    3.34e+06
```

```
==========================================================================
===
Ljung-Box (L1) (Q):                   3.39   Jarque-Bera (JB):               79
3.12
Prob(Q):                              0.07   Prob(JB):
0.00
Heteroskedasticity (H):              15.62   Skew:
0.00
Prob(H) (two-sided):                  0.00   Kurtosis:                        1
2.16
==========================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 33088.58251747072
                               SARIMAX Results
==========================================================================
Dep. Variable:            Chicago, IL   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -1863.466
Date:                Fri, 13 May 2022   AIC                         3736.932
Time:                        12:06:37   BIC                         3754.056
Sample:                    04-01-1996   HQIC                        3743.842
                         - 04-01-2015
Covariance Type:                 opg
==========================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------
ar.L1          0.9237      0.204      4.538      0.000       0.525       1.323
ma.L1         -0.9113      0.203     -4.493      0.000      -1.309      -0.514
ma.L2         -0.0235      0.023     -1.045      0.296      -0.068       0.021
ma.L3         -0.0037      0.027     -0.136      0.892      -0.056       0.049
sigma2      7.277e+05   2.76e+04     26.409      0.000    6.74e+05    7.82e+05
==========================================================================
===
Ljung-Box (L1) (Q):                   9.66   Jarque-Bera (JB):               79
5.02
Prob(Q):                              0.00   Prob(JB):
0.00
Heteroskedasticity (H):              26.99   Skew:
0.56
Prob(H) (two-sided):                  0.00   Kurtosis:                        1
2.10
==========================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 3961.5119734358436
                               SARIMAX Results
==========================================================================
Dep. Variable:          Baltimore, MD   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -1746.322
Date:                Fri, 13 May 2022   AIC                         3502.644
Time:                        12:06:37   BIC                         3519.768
Sample:                    04-01-1996   HQIC                        3509.554
                         - 04-01-2015
Covariance Type:                 opg
==========================================================================
```

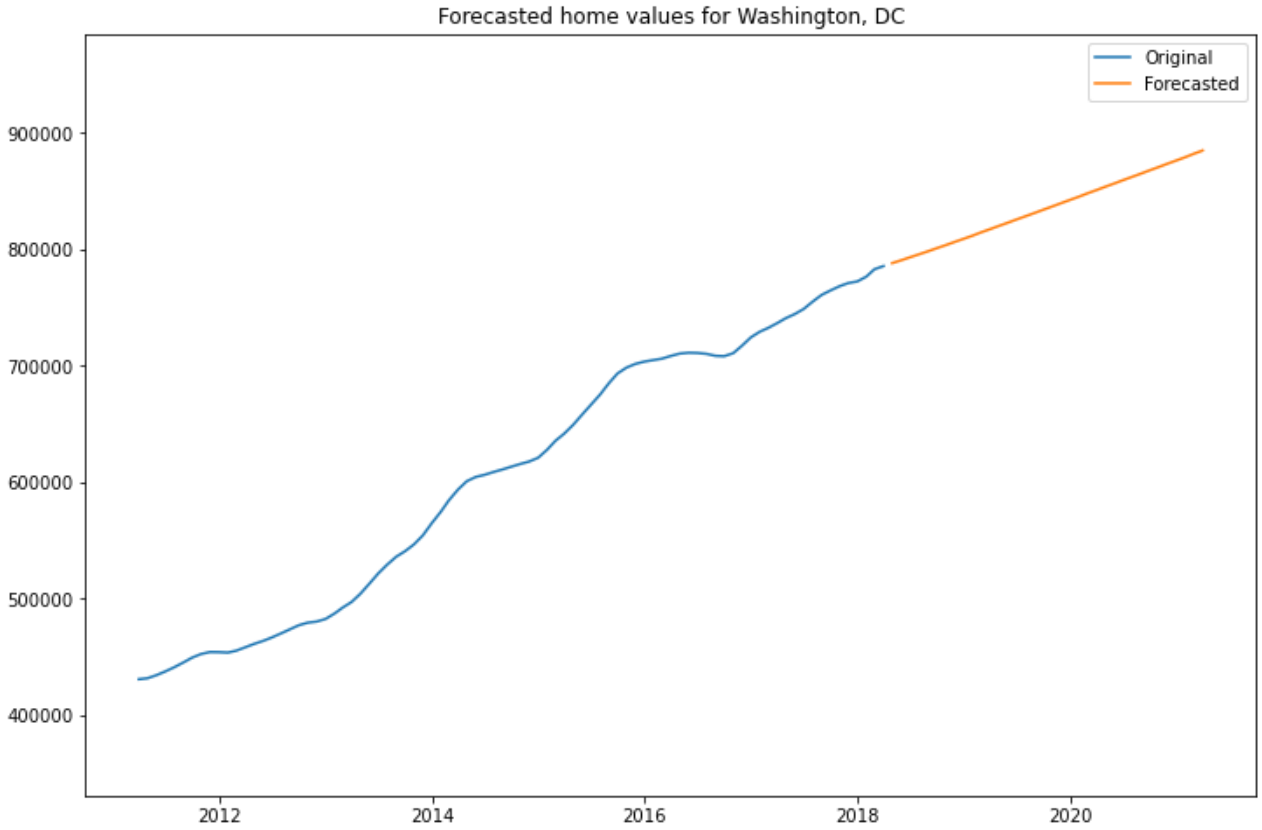
```
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4565      0.464      0.984      0.325      -0.453       1.366
ma.L1         -0.4302      0.460     -0.935      0.350      -1.332       0.472
ma.L2         -0.1856      0.028     -6.664      0.000      -0.240      -0.131
ma.L3          0.0214      0.092      0.233      0.816      -0.159       0.202
sigma2      2.556e+05    1.2e+04     21.356      0.000    2.32e+05    2.79e+05
===================================================================================
===
Ljung-Box (L1) (Q):                   2.19   Jarque-Bera (JB):                49
0.25
Prob(Q):                              0.14   Prob(JB):
0.00
Heteroskedasticity (H):              13.02   Skew:
1.13
Prob(H) (two-sided):                  0.00   Kurtosis:
9.83
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 16300.327545138
                               SARIMAX Results
==============================================================================
Dep. Variable:             Boston, MA   No. Observations:                  229
Model:                 ARIMA(1, 2, 3)   Log Likelihood               -1993.505
Date:                Fri, 13 May 2022   AIC                         3997.010
Time:                        12:06:38   BIC                         4014.134
Sample:                    04-01-1996   HQIC                        4003.920
                         - 04-01-2015
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7180      0.135      5.335      0.000       0.454       0.982
ma.L1         -0.8022      0.136     -5.894      0.000      -1.069      -0.535
ma.L2         -0.0284      0.040     -0.714      0.475      -0.106       0.050
ma.L3         -0.0250      0.027     -0.935      0.350      -0.078       0.027
sigma2      2.273e+06   7.64e+04     29.768      0.000    2.12e+06    2.42e+06
===================================================================================
===
Ljung-Box (L1) (Q):                   2.94   Jarque-Bera (JB):             1074
6.78
Prob(Q):                              0.09   Prob(JB):
0.00
Heteroskedasticity (H):              19.02   Skew:                           -
1.47
Prob(H) (two-sided):                  0.00   Kurtosis:                        3
6.58
===================================================================================
===

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-s
tep).
------------------------
------------------------
RMSE: 53421.36806305324
```

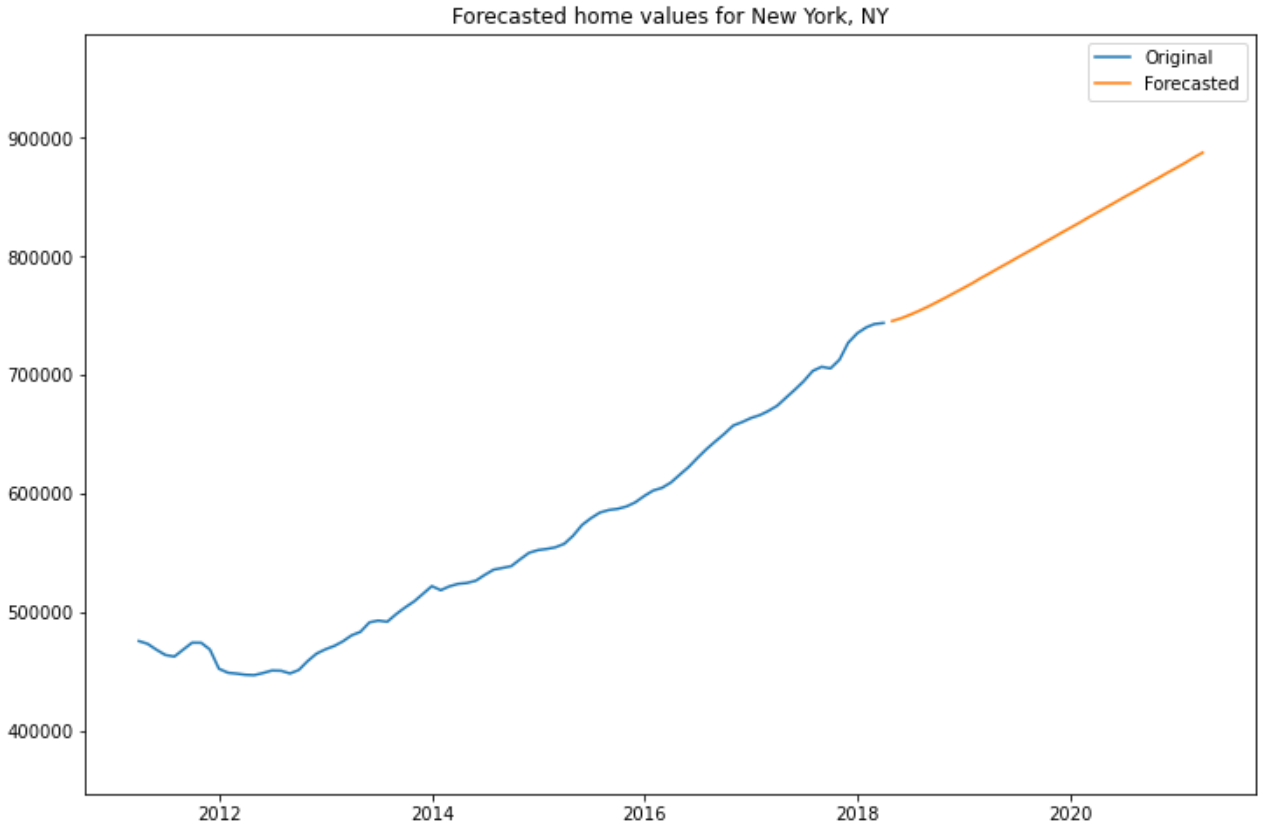


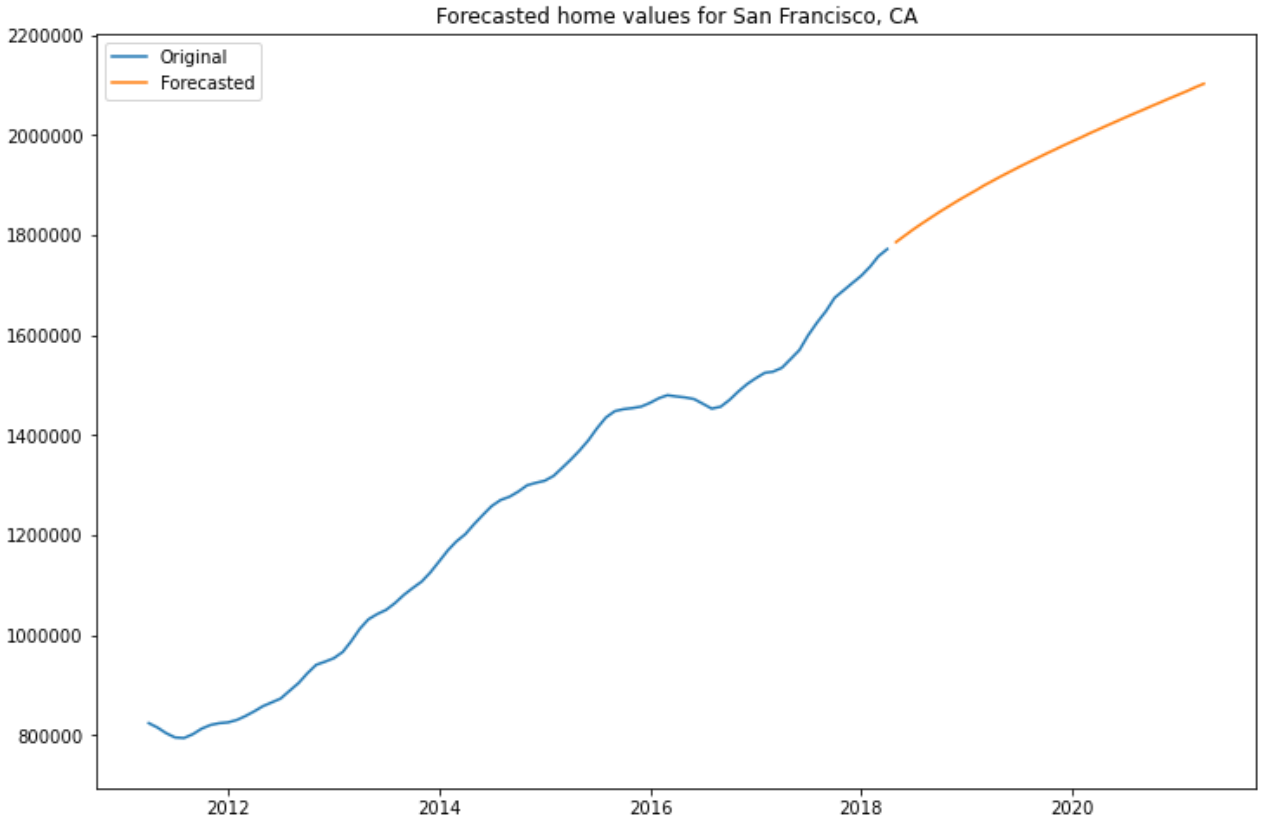
Forcasting home price with best model

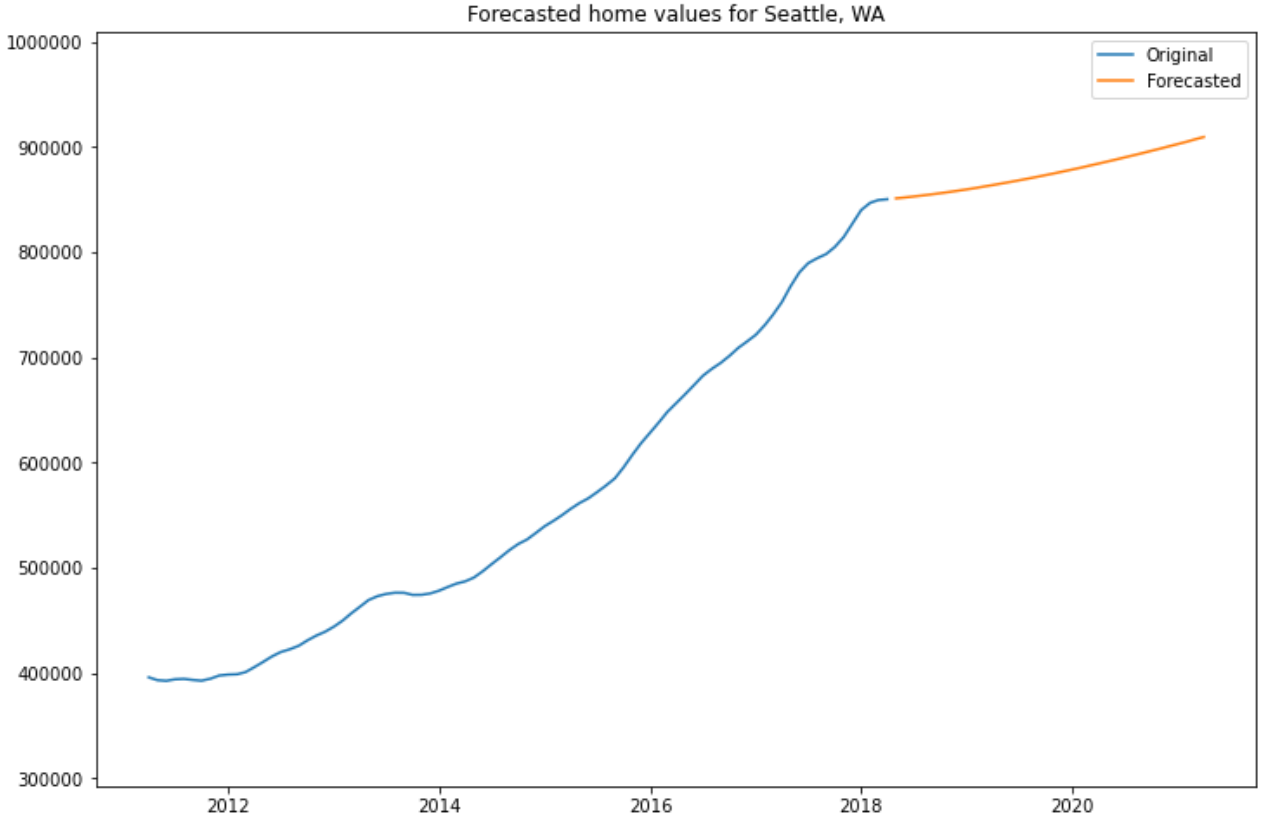
In [94]:
```
#We created an empty dictionary for our predictions
predictions_dict = {}

#We are running a for-loop where we add the forecast to the dictionary for every
for city in city_list:
    forecast_mod = ARIMA(melted_df[city], order = [1,2,3]).fit()
    predictions_dict[city] = forecast_mod.forecast(steps=36)
```

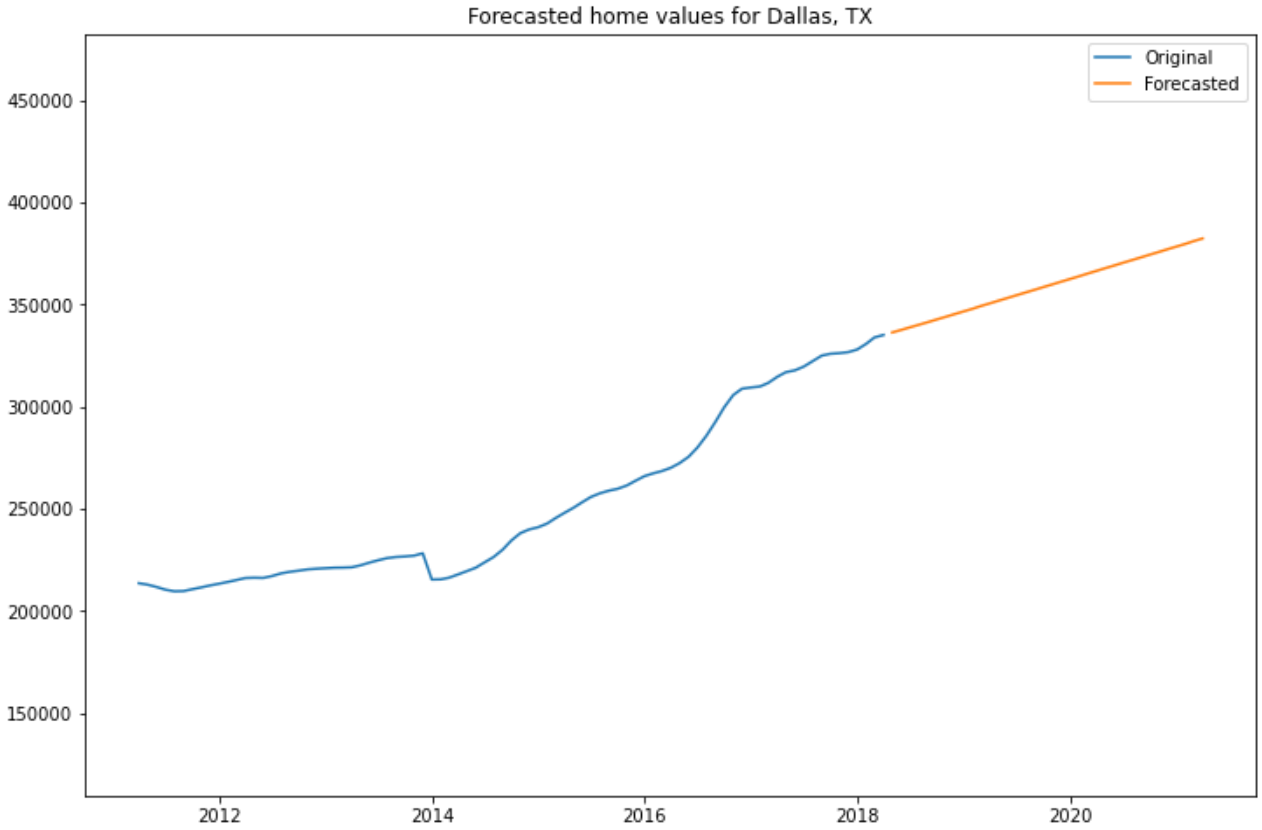
In [96]:
```
#We made a for loop that graphs all of the predictions
for city, predictions in predictions_dict.items():
    fig, ax = plt.subplots(figsize=(12,8))
    ax.plot(melted_df[180:][city])
    ax.plot(predictions)
    ax.legend(['Original', 'Forecasted'])
    ax.set_title(f'Forecasted home values for {city}')
    ax.ticklabel_format(axis='y',style='plain')
    ax.set_ylim([min(melted_df[180:][city]-100000),max(predictions)+100000])
    plt.show()
```

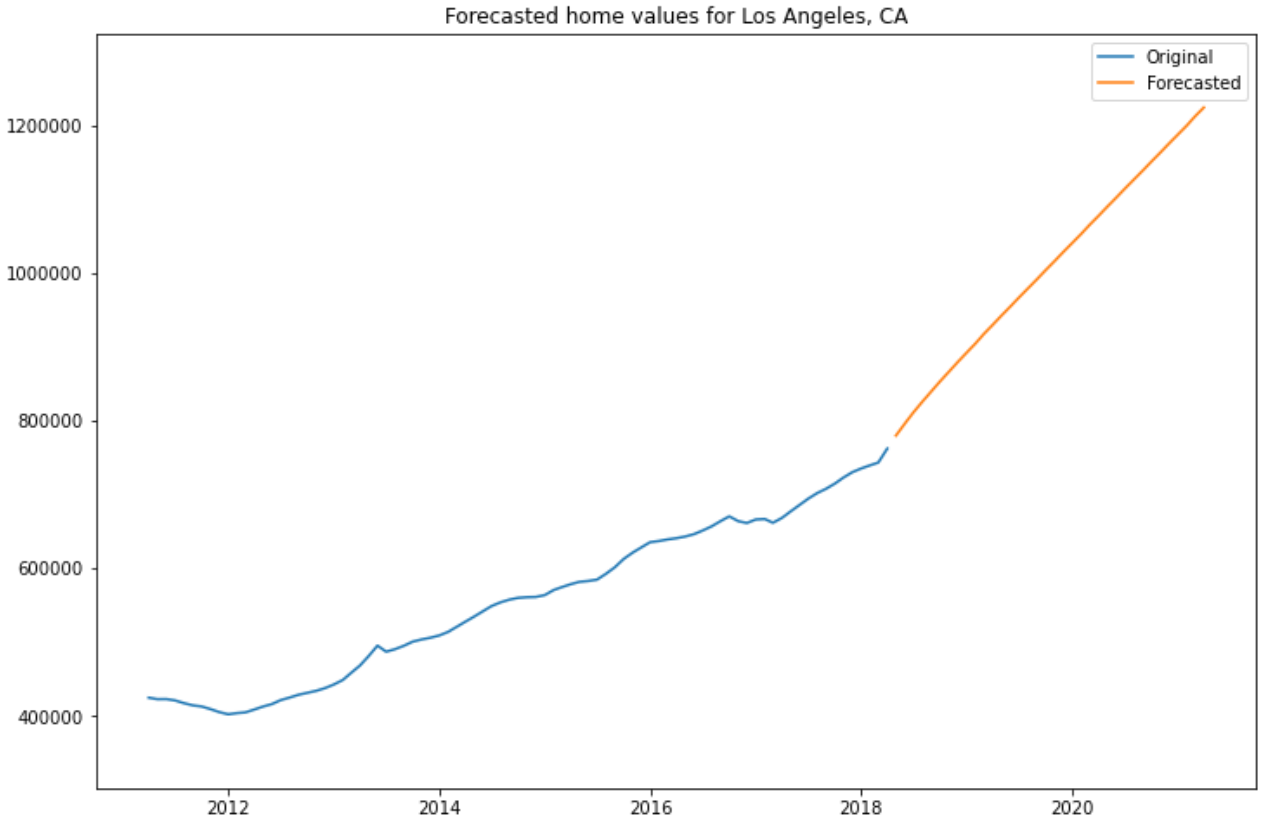

Forecasted home values for Washington, DC
Original
Forecasted
900000
800000
700000
600000
500000
400000
2012
2014
2016
2018
2020

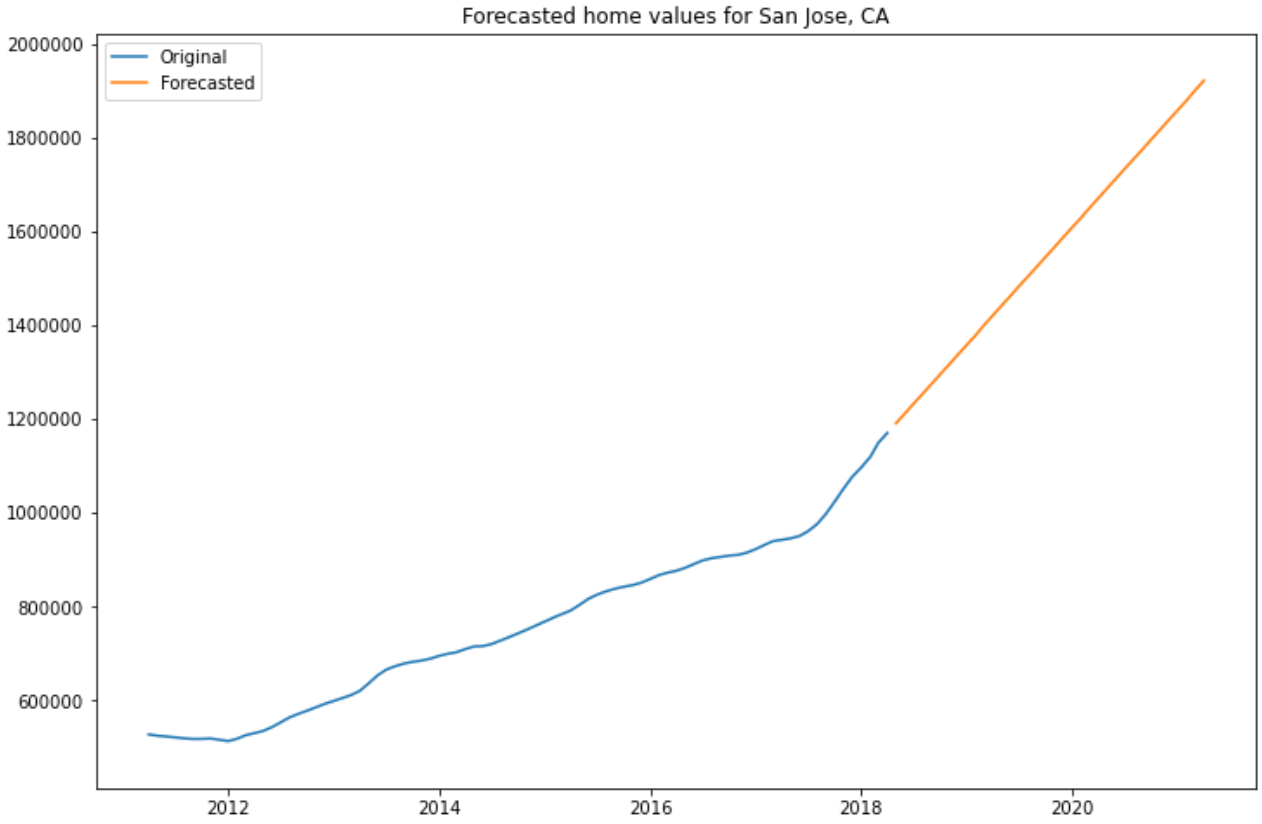

Forecasted home values for New York, NY
Original
Forecasted
900000
800000
700000
600000
500000
400000
2012
2014
2016
2018
2020

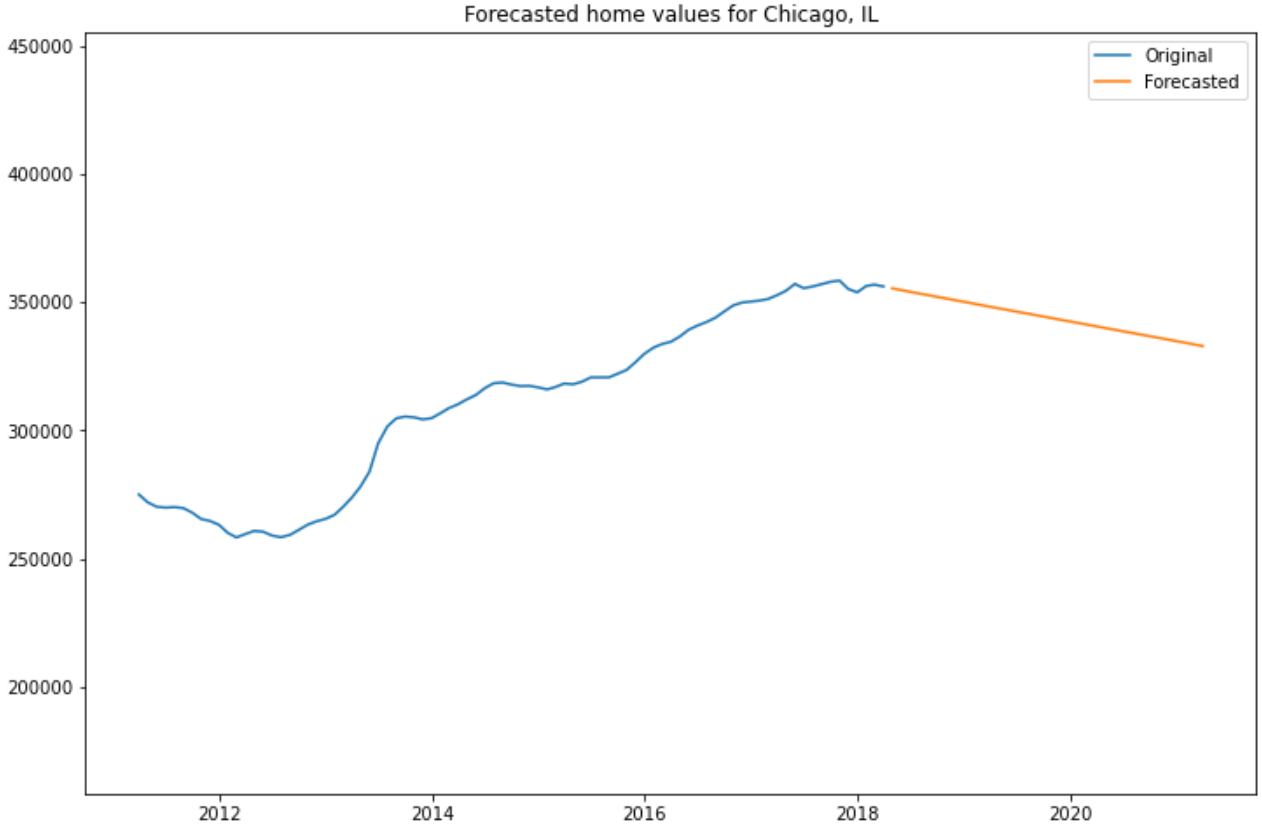

Forecasted home values for San Francisco, CA
Original
Forecasted
2200000
2000000
1800000
1600000
1400000
1200000
1000000
800000
2012
2014
2016
2018
2020

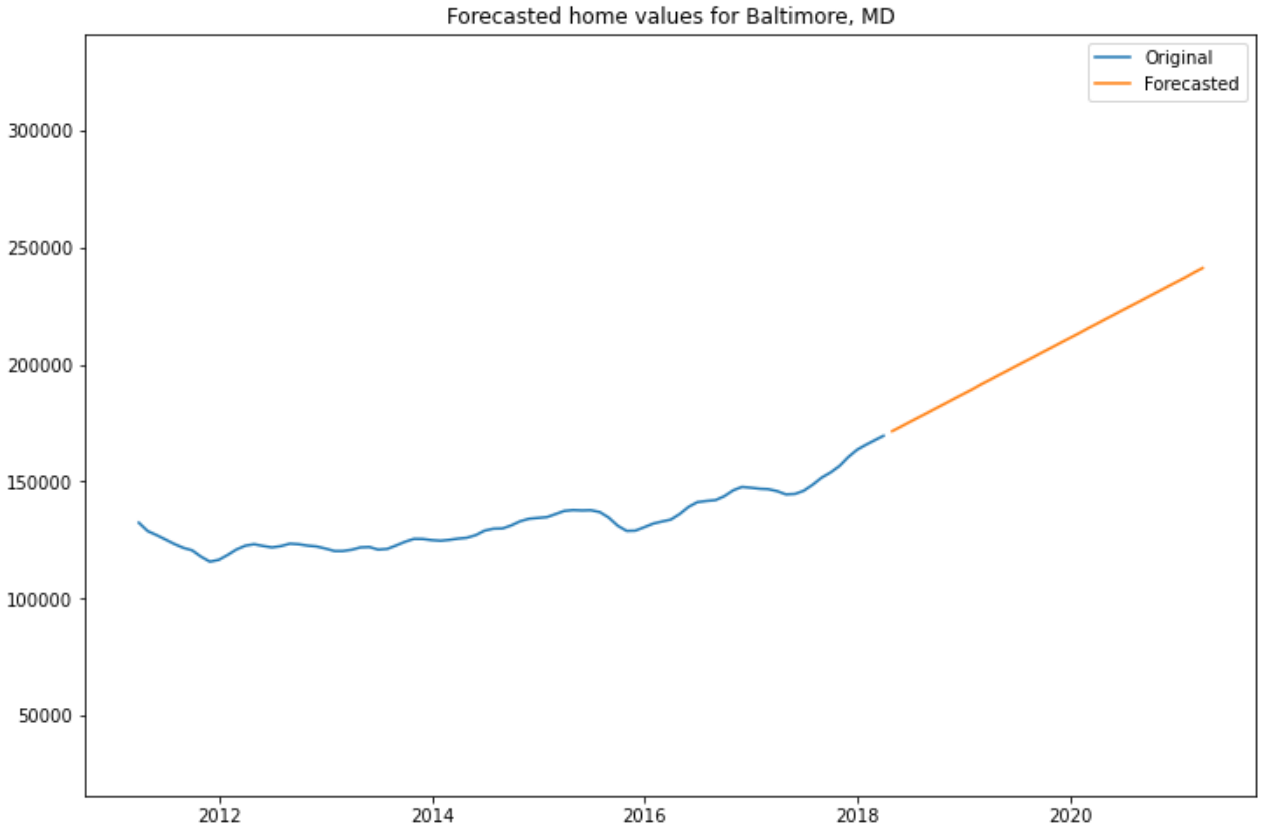

Forecasted home values for Seattle, WA
Original
Forecasted
1000000
900000
800000
700000
600000
500000
400000
300000
2012
2014
2016
2018
2020

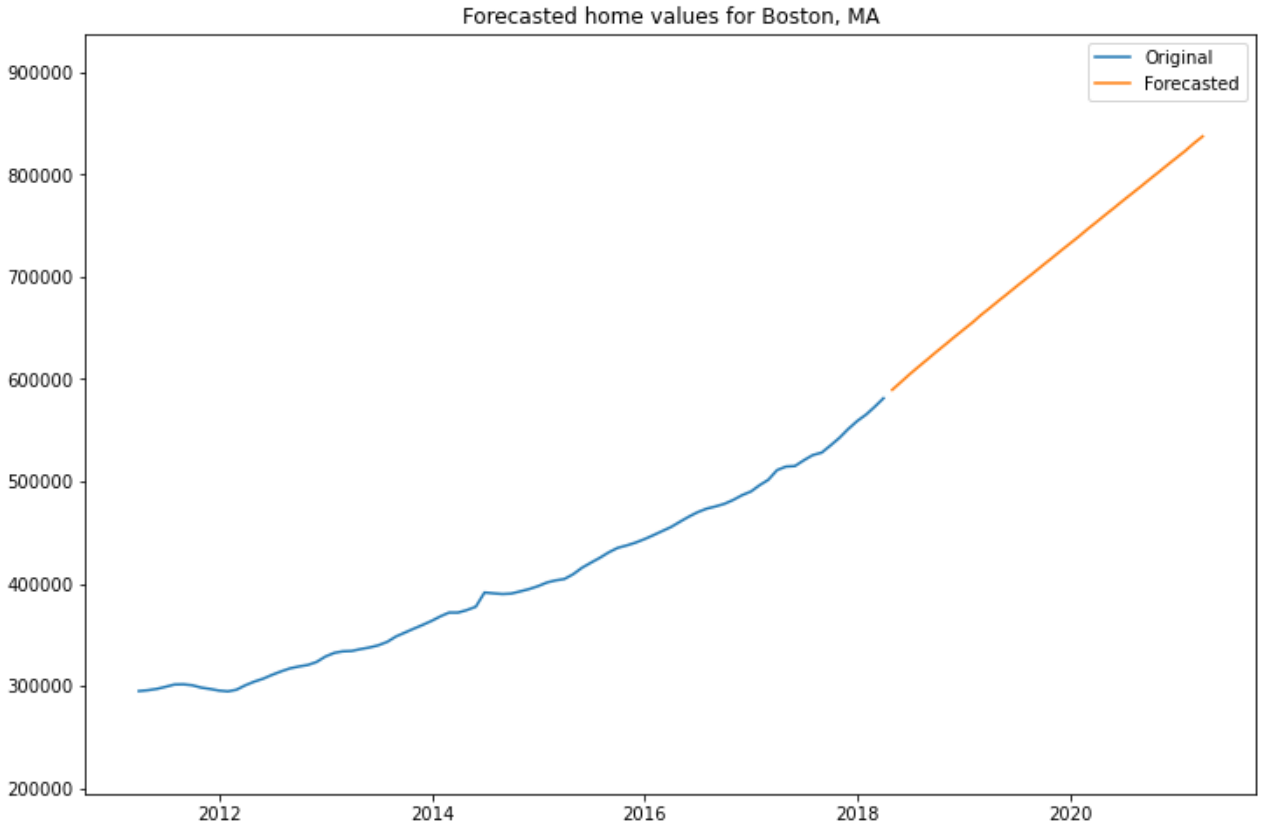

Forecasted home values for Dallas, TX
Original
Forecasted
450000
400000
350000
300000
250000
200000
150000
2012
2014
2016
2018
2020


Forecasted home values for Los Angeles, CA
Original
Forecasted
1200000
1000000
800000
600000
400000
2012
2014
2016
2018
2020


Forecasted home values for San Jose, CA
Original
Forecasted
2000000
1800000
1600000
1400000
1200000
1000000
800000
600000
2012
2014
2016
2018
2020


Forecasted home values for Chicago, IL
Original
Forecasted
450000
400000
350000
300000
250000
200000
2012
2014
2016
2018
2020





In [97]:
```
forcasted_return = {}
for city in city_list:
    forcast_return = ((predictions_dict[city][-1]/melted_df[city][-1])*100) -100
    forcasted_return[city]=forcast_return

forcasted_return = pd.DataFrame.from_dict(forcasted_return, orient='index')
forcasted_return.reset_index(inplace=True)
```

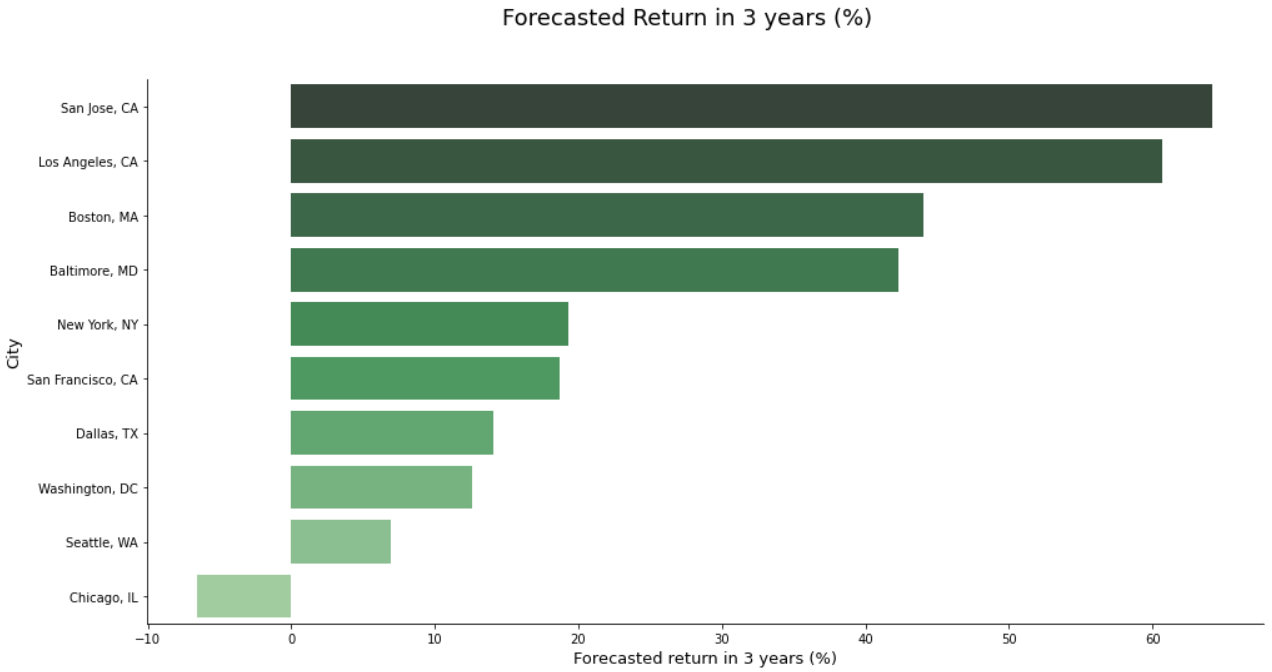
```
forcasted_return.rename({0:'Percent Return in 3 years','index':'City'},inplace=T
forcasted_return
```

Out[97]:

	City	Percent Return in 3 years
0	Washington, DC	12.635529
1	New York, NY	19.307380
2	San Francisco, CA	18.655403
3	Seattle, WA	6.953193
4	Dallas, TX	14.096486
5	Los Angeles, CA	60.669809
6	San Jose, CA	64.161839
7	Chicago, IL	-6.531644
8	Baltimore, MD	42.280752
9	Boston, MA	44.026092

In [98]:

```
fig, ax = plt.subplots(figsize=(16,8))
forcasted_sort = forcasted_return.sort_values('Percent Return in 3 years',ascend
# sns.set_palette("crest")

pal = sns.color_palette("Greens_d", len(forcasted_sort))
# rank = forcasted_sort.argsort().argsort()   # http://stackoverflow.com/a/62665
sns.barplot(x='Percent Return in 3 years', y='City', data=forcasted_sort,palette
sns.despine()
fig.suptitle("Forecasted Return in 3 years (%)", fontsize = 18)
ax.set_xlabel("Forecasted return in 3 years (%)", fontsize = 13)
ax.set_ylabel("City", fontsize = 13)
plt.savefig('figures/forecast_return.png', transparent=True,bbox_inches="tight",
plt.savefig('figures/forecast_return.jpeg', transparent=True,bbox_inches="tight"
```



Conclusions

Recommendations

The forecasts show the largest growth in San Jose, Los Angeles, Boston, Baltimore, and New York. We would suggest focusing on these markets. San Francisco was a close runner-up, but the market is much more expensive there. We would be unable to buy as much property there, and would be able to diversify much more in cities like Baltimore or Los Angeles. There is a good balance to the 5 cities we have suggested here that should hedge against itself. San Jose, Boston, and New York are well established cities with not a lot of buildable land left. Just owning real estate in these cities will ensure that values will give consistent returns. However, Baltimore and Los Angeles have more land that could give huge returns if invested in and renovated properly. The trick would be doing this in a manner that does not feel to be undermining the affordability, culture, or diversity of these neighborhoods as is often the case. These rennovations must not feel gaudy, but seemless and at home with the current residents. Another more ethically straightforward, but politically challenging idea is looking into building more dense housing in the suburbs with multiplexes or townhouses. These projects often run into obstacles, but as NIMBY culture becomes less popular and zoning reform progresses, new investment opportunities should arise.

Next Steps

There are several steps that could be taken to give even more value if we had the funding. While our model performed quite well, it was evaluated on test data that was consistently a bull market. If the market was bear or particularly volatile, it is unlikely that the model would perform as well. This is particularly seen with the prediction for Chicago market. Because Chicago was in a short term slump in our most recent data points, this market shrinkage was projected to continue for the next three years. Designing a model that could differentiate volatility changes as opposed to structural market failures could allow for much better predictions.

Furthermore, our model only looked at factors endogenous to the time series data. While this is useful for understanding how real estate markets change over time, it does nothing to explain all the other factors that are driving changes in housing prices. If we had more data on factors relevant to housing prices, such as housing density, quality of infrastructure, and cultural engagement, then we could explain so much more of the variance that our model failed to explain. In particular, we could use more complex models like SARIMAX, or even a Long-short Term Memory Neural Network to catch on to patterns completely unexplored by our current model.

Finally, since this data has been recorded, much has changed in the real estate market. COVID-19 made the real estate market come to an abrupt halt, only for the absurdly low interest rates to trigger one of the greatest housing market shortages in decades. And now, with the interest rates increasing once more, it seems that the housing market is starting to cool off once more. A simple ARIMA model like ours would be completely inadequate to analyzing all the crazy

changes seen over the last three years. Having more recent data could give very useful insights on understanding many different phenomena induced by the pandemic.