Battle of Neighborhoods - Auberge Beau Sejour

Coursera Capstone Report

Yves Bourgeois

Table of contents¶

Part 1 – Project Outline

Part 2 - Working with Data

Chapter 13 - SARIMAX Forecasting Future Median Sale Prices¶

Chapter 14 - Competitive Analysis¶

Chapter 15 - Comparison of Properties on MLS¶

Chapter 16 - Notes and Sources and Methodology¶

Part 1 – Project Outline¶

i A description of the problem and a discussion of the background

ii Background

Over the last few years there has been an explosion of AIR-BnB properties popping up across the state of California with no signs of abating. There is a wide variety of properties from bachelor pads to 8 room mansions. However, there remains a niche market that AirBnB does not compete with directly, namely Bed & Breakfasts. Some travelers prefer the rustic charm of a smaller place to stay and have the opportunity to chat with the owner and locals. They are looking to absorb the local culture and enjoy the outdoors free from unfettered commercialism like Disneyland or SeaWorld.¶

My friend is an architect and has travelled the world throughout his career and is soon to retire in southern California. He is looking to focus on this niche market and open a luxury Bed & Breakfast and run it as a family business. This will allow him to meet lots of interesting people without having to travel himself. His vision is to open a retreat that focuses on nature and tranquility. The B&B will offer yoga classes, massage therapies, and organic cuisine. He plans on calling it *Auberge Beau Sejour*¶

He made a lot of money when his company recently went IPO. Therefore, his budget is 4 million dollars to buy a property and convert it into a B&B. Real estate along the California coast is very expensive, and 4 million dollars is likely the minimum amount needed to buy something suitable near the ocean. The availability of properties is scarce, and this project might not be a success. My friend still travels a lot, and does not have the time to research multiple web sites and find the necessary information. He will offer me a free two week stay, if I complete this project for him. Nevertheless, he has a good idea of what he is looking for. Here is a summary of the minimum requirements he proposes.¶

Tourism venues

- The property must be close to several outdoor venues like beaches and parks.
- The property's must be close to venues like wine tasting rooms and wineries.

Demographic statistics

- The chosen county must have a crime rate well below average for the state of California.
- The chosen city must have at least 30,000 per capita income.

Financial criteria

- The total cost of the property must be below 4 million dollars.¶

Geographic markers

- The property must be within a 2-hour drive of LAX airport.
- The property must be close as possible to highway 101, with access to the beach having a higher priority.

Property characteristics

- The property should be able to accommodate a minimum of six guests or have at least four bedrooms.¶

iii Defining the Business Problem

The business problem is centered around finding a piece of real estate that fits all of these conditions.¶

iv Data Sources

The key data sources for this project are as follows:

1- Redfin and Trulia and Zillow for MLS listings, market analysis and forecasting.

2- Wikipedia for demographic information

3- Foursquare for venue information by city and competitive intelligence

4- Google maps showing driving distances between cities and airports. Folium for interactive maps showing the city locations based on criteria like PCI and crime rates.

v Methodology

The methodology used will be explained throughout chapters 11 through chapter 13. Chapter 11 covers historical data for median sale price trending. Chapter 12 cover various forecasting models such as AR, MA and ARIMA. Chapter 13 will cover the SARIMAX forecasting model.

vi Observations / Discussion

The symbol of a lightbulb will appear when observations or insights are made about the methodology or source data throughout this report.

vii Results

The results from various statistical tests will be explained throughout chapters 11 through chapter 13. Chapter 11 covers historical data for median sale price trending. Chapter 12 cover various forecasting models such as AR, MA and ARIMA. Chapter 13 will cover the SARIMAX forecasting model.

PART 2 – Working with Data

A change in focus

In Part 1 of this project, one analysis step was to conduct clustering using k-means. However, after analyzing the cities with Foursquare, it became clear that clustering venues would not be helpful, due to the low number of venues for each city. Furthermore, as we will see later in chapter 14, there are too few competitors to do any clustering. In addition, access to MLS listings is restricted and clustering of available properties is not possible. Nevertheless, there are plenty of criteria to reduce the number of potential candidates, such as crime rates, per capita income, driving times to LAX airport, proximity to the beach and parks to name a few. Therefore, it was decided to focus on two areas, the first was mapping using these criteria and the second was using inferential statistical testing to predict future median sale prices with Python.

1. Create Notebook and download libraries and packages

There are too many cities in California to research in a timely manner. The goal is to reduce the number of cities as quickly as possible. Therefore, we begin our search further up the hierarchy at the county level. After reducing the number of candidates at the top of the funnel, we can move our way down to the city level.

1.a) Evaluate the crime rate by county in the state of California

The owner of the Bed & Breakfast is looking to build a peaceful oasis away from commercialism and crime. This retreat must attract locals with a good income as well as international travelers arriving from LAX airport. Therefore, a key criterion to this search is looking at crime rates in California. Crime rates vary dramatically by region and category. "The lowest rates of both violent and property crime in 2017 were on the South Coast (Imperial, Orange, San Diego, and Ventura Counties), with rates of 288 and 1,894 per 100,000 residents, respectively."[1] Based on this fact, we can narrow the search from fifty-eight counties down to four counties in Southern California.

Source: Public Policy Institute of California https://www.ppic.org/publication/crime-trends-in-california/

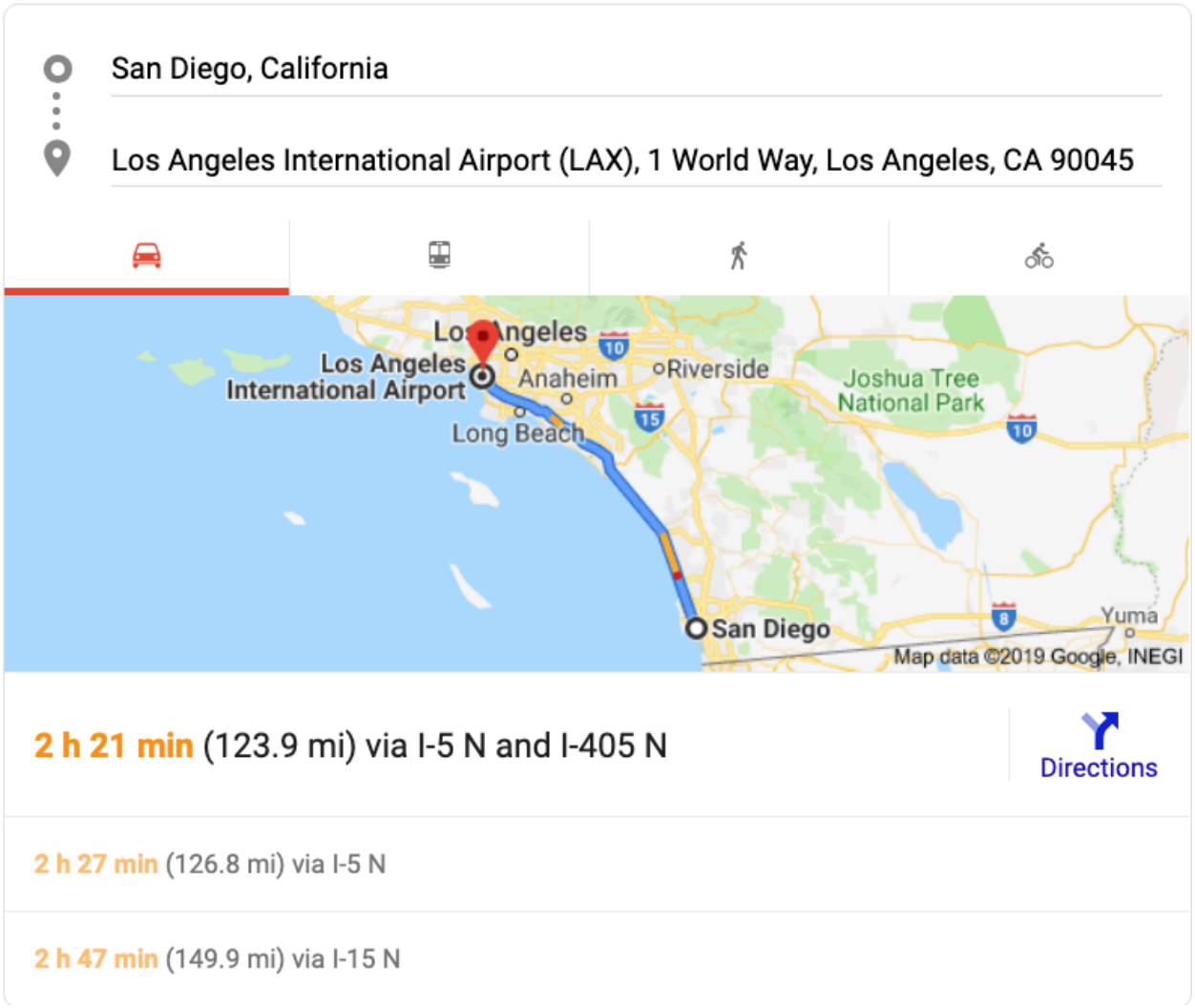
1.b) Identify top 4 counties with the lowest crime rate in the state of California

Based on this information on crime rates in California, the list of possible locations is now reduced to four, namely 1- Imperial County, 2- San Diego County, 3- Ventura County and 4- Orange County

These counties are relatively close to LAX airport in Los Angeles county, a key search criteria for this project. Another key search criterion outlined in the introduction is that the Bed & Breakfast should be relatively close to the ocean. Looking at the counties in yellow on the map, we see that Imperial County would likely lose out to the other counties because it is not close enough to the ocean. Therefore, we can eliminate it as a viable option.

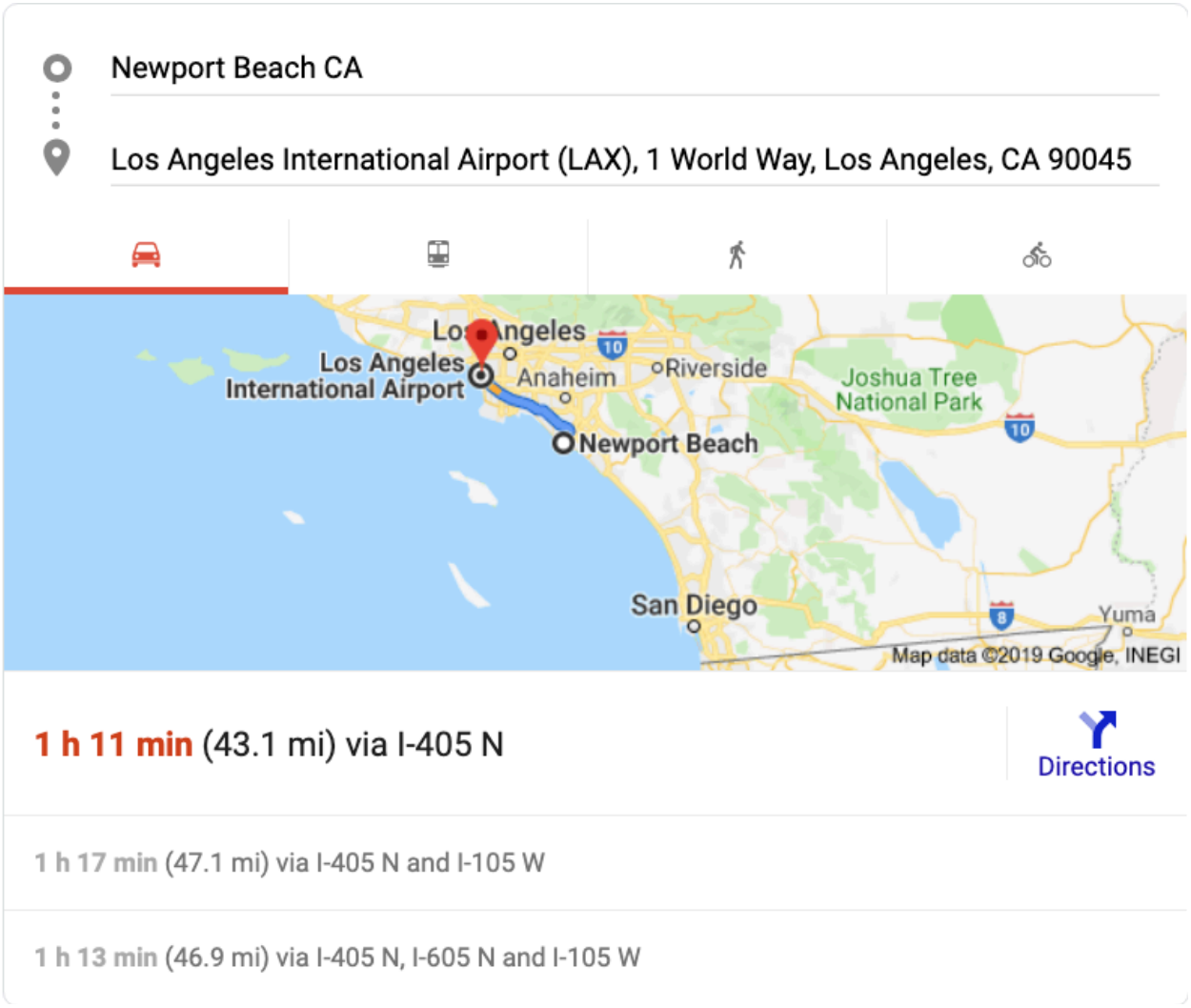
2. The next step in narrowing the focus is to look at driving times and distances from LAX¶

2.a) The county furthest from LAX is San Diego



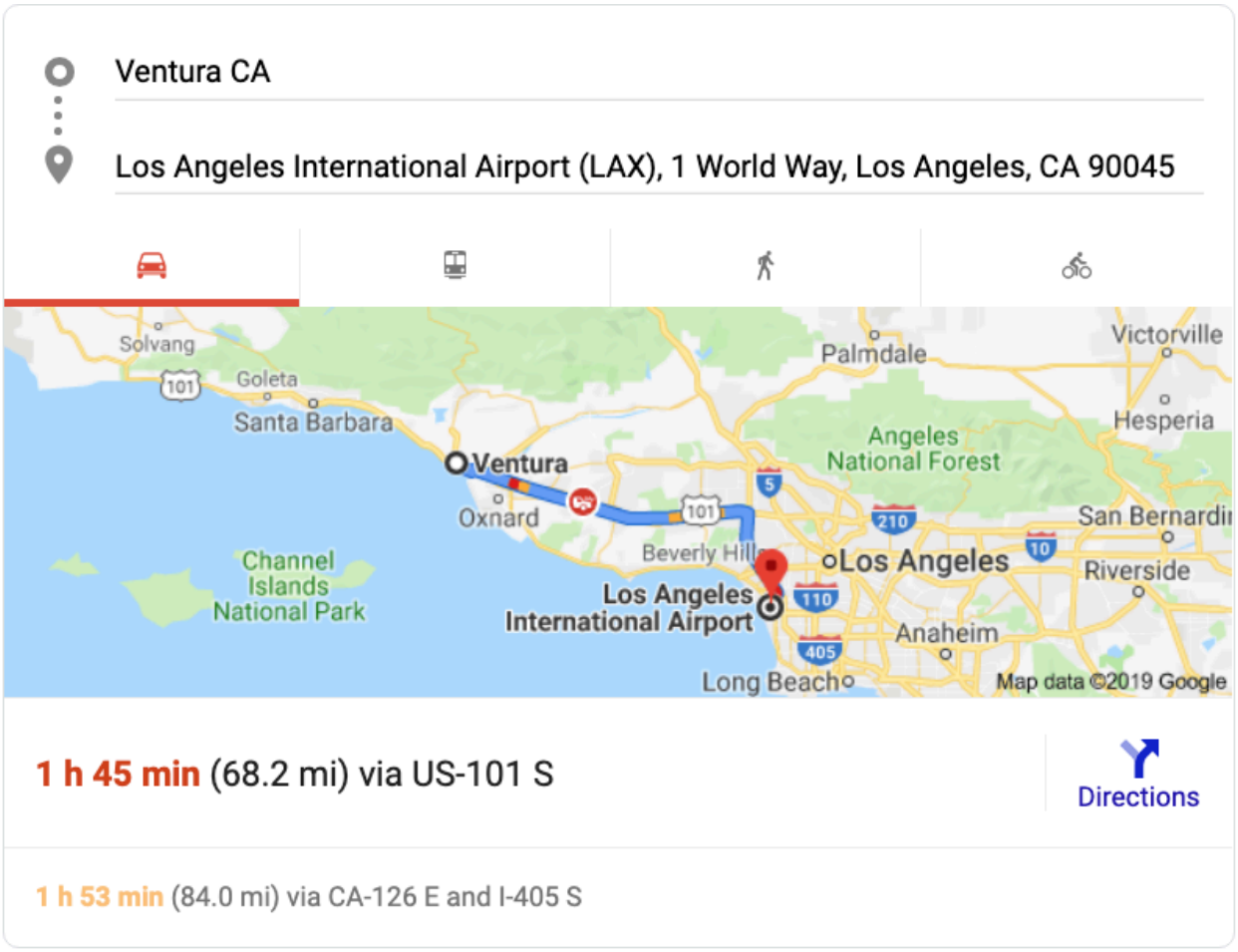
The driving time is greater than 2 hours and is no longer a viable option.

2.b) The next option is Orange County¶



Orange County remains a viable option because the drive time to LAX is less than 2 hours.

2.c) Ventura County remains the last option to evaluate for distance and driving times¶



Ventura County remains a viable option because the drive time to LAX is less than 2 hours. In summary, after reviewing the driving distance to the LAX airport, we are left with two viable options, namely Ventura and Orange counties. The next step is to look at the crime rate by city in each county.

3. Evaluate the crime rate of cities in each county¶

A quick search on Wikipedia, there are approximately 20 cities and towns in Ventura county and about 34 in Orange county. Therefore, we need to shorten the list of cities by using search criteria like crime rates and per capita income.

3.a) Evaluate the crime rate of cities in Ventura county¶

Out[2]:

	City	Population	Violent crimes	Violent crime rateper 1,000 persons	Property crimes	Property crime rateper 1,000 persons
0	Camarillo	66506	61	0.92	955	14.36
7	Simi Valley	126686	141	1.11	1916	15.12
2	Moorpark	35102	41	1.17	330	9.40
8	Thousand Oaks	129171	157	1.22	1838	14.23
1	Fillmore	15298	24	1.57	198	12.94
3	Ojai	7607	13	1.71	162	21.30
9	Ventura	108511	310	2.86	3885	35.80
5	Port Hueneme	22142	65	2.94	467	21.09
4	Oxnard	201797	603	2.99	4071	20.17
6	Santa Paula	29899	91	3.04	590	19.73

3.b) Evaluate the crime rate of cities in Orange County¶

	City	Population	Violent crimes	Violent crime rateper 1,000 persons	Property crimes	Property crime rateper 1,000 persons
15	Laguna Woods	16590	4	0.24	148	8.92
31	Villa Park	5956	2	0.34	87	14.61
11	Irvine	217528	110	0.51	3304	15.19
24	Rancho Santa Margarita	49038	27	0.55	319	6.51
28	Seal Beach	24764	17	0.69	545	22.01
14	Laguna Niguel	64533	47	0.73	764	11.84
20	Mission Viejo	95599	73	0.76	1197	12.52
33	Yorba Linda	65820	53	0.81	787	11.96
0	Aliso Viejo	48999	43	0.88	415	8.47
13	Laguna Hills	31090	29	0.93	620	19.94
22	Orange	139692	135	0.97	2833	20.28
18	La Palma	15954	18	1.13	340	21.31
5	Cypress	48976	56	1.14	1018	20.79
25	San Clemente	65089	75	1.15	839	12.89
21	Newport Beach	87286	101	1.16	2151	24.64
17	Lake Forest	79166	107	1.35	1088	13.74
30	Tustin	77400	114	1.47	1653	21.36
10	Huntington Beach	194677	313	1.61	5470	28.10
26	San Juan Capistrano	35449	59	1.66	519	14.64
2	Brea	40253	74	1.84	1292	32.10
7	Fountain Valley	56674	106	1.87	1469	25.92
6	Dana Point	34172	65	1.90	604	17.68
23	Placentia	51778	107	2.07	906	17.50
4	Costa Mesa	112635	254	2.26	4079	36.21
19	Los Alamitos	11728	27	2.30	357	30.44
16	La Habra	61731	147	2.38	1150	18.63
12	Laguna Beach	23283	57	2.45	548	23.54
3	Buena Park	82505	206	2.50	2066	25.04
9	Garden Grove	175079	439	2.51	4017	22.94
29	Stanton	39124	104	2.66	630	16.10

3.c) Determine cities with the lowest violent crime rates of both counties (54 cities total)¶

City	Population	Violent crimes	Violent crime rateper 1,000 persons	Property crimes	Property crime rateper 1,000 persons
Laguna Woods	16590	4	0.24	148	8.92
Villa Park	5956	2	0.34	87	14.61
Rancho Santa Margarita	49038	27	0.55	319	6.51
Seal Beach	24764	17	0.69	545	22.01
Laguna Niguel	64533	47	0.73	764	11.84
Mission Viejo	95599	73	0.76	1197	12.52
Yorba Linda	65820	53	0.81	787	11.96
Aliso Viejo	48999	43	0.88	415	8.47
Camarillo	66506	61	0.92	955	14.36
Laguna Hills	31090	29	0.93	620	19.94
La Palma	15954	18	1.13	340	21.31
Cypress	48976	56	1.14	1018	20.79
San Clemente	65089	75	1.15	839	12.89
Moorpark	35102	41	1.17	330	9.40
Thousand Oaks	129171	157	1.22	1838	14.23
Lake Forest	79166	107	1.35	1088	13.74
Tustin	77400	114	1.47	1653	21.36
Fillmore	15298	24	1.57	198	12.94
San Juan Capistrano	35449	59	1.66	519	14.64
Ojai	7607	13	1.71	162	21.30
Brea	40253	74	1.84	1292	32.10
Fountain Valley	56674	106	1.87	1469	25.92
Dana Point	34172	65	1.90	604	17.68
Placentia	51778	107	2.07	906	17.50
Los Alamitos	11728	27	2.30	357	30.44
La Habra	61731	147	2.38	1150	18.63
Laguna Beach	23283	57	2.45	548	23.54
Stanton	39124	104	2.66	630	16.10

Removed Cities above violent and property crime thresholds

4.a) Evaluate the per capita income of each city in both counties¶

Ventura County

Based on selection criteria we apply a $30k threshold to per capita income. San Buenaventura (Ventura) appears to be the last city on the list that fits this criterion

	Place	Type	Population	Per capita income	Median household income	Median family income
0	Bell Canyon	CDP	2291	$85,789	$220,764	$230,455
18	Santa Rosa Valley	CDP	3143	$71,594	$154,931	$176,938
10	Oak Park	CDP	14045	$55,681	$128,618	$143,188
3	Channel Islands Beach	CDP	3299	$53,891	$70,313	$69,333
22	Thousand Oaks	City	125633	$46,093	$100,373	$112,876
19	Santa Susana	CDP	1115	$40,271	$111,610	$112,027
11	Oak View	CDP	4166	$38,062	$80,614	$81,750

	Place	Type	Population	Per capita income	Median household income	Median family income
1	Camarillo	City	64340	$37,840	$84,168	$101,334
12	Ojai	City	7496	$36,769	$63,750	$89,338
9	Moorpark	City	34100	$36,375	$103,009	$107,412
21	Simi Valley	City	122864	$35,467	$89,452	$97,999
7	Meiners Oaks	CDP	3339	$33,264	$51,955	$73,365
8	Mira Monte	CDP	7666	$32,718	$71,723	$83,968
16	San Buenaventura (Ventura)	City	105809	$31,775	$66,226	$81,616
2	Casa Conejo	CDP	3424	$26,950	$84,286	$86,630

Orange County

Based on selection criteria we apply a $30k threshold to per capita income. Placentia appears to be the last city on the list that fits this criterion.

	Place	Type	Population	Per capita income	Median household income	Median family income
15	Laguna Beach	City	22808	$81,591	$99,190	$139,833
26	Newport Beach	City	84417	$80,872	$108,946	$151,773
39	Villa Park	City	5825	$71,697	$151,139	$165,833
6	Coto de Caza	CDP	14974	$65,625	$164,385	$176,686
27	North Tustin	CDP	24572	$55,038	$109,629	$119,543
2	Anaheim Hills	City	55036	$52,195	$123,260	$148,360
17	Laguna Niguel	City	62855	$51,491	$100,480	$119,757
8	Dana Point	City	33510	$51,431	$83,306	$101,186
31	Rossmoor	CDP	10099	$51,210	$108,427	$119,727
41	Yorba Linda	City	63578	$49,485	$115,291	$128,528
14	Ladera Ranch	CDP	21412	$48,671	$132,475	$143,857
32	San Clemente	City	62052	$47,894	$89,289	$107,524
37	Sunset Beach	CDP	1486	$47,415	$68,036	$109,125
22	Las Flores	CDP	5911	$46,717	$128,269	$135,046
16	Laguna Hills	City	30477	$44,751	$85,971	$105,385
0	Aliso Viejo	City	47037	$44,646	$99,095	$113,183
35	Seal Beach	City	24157	$44,115	$50,958	$94,035
13	Irvine	City	205057	$43,102	$92,599	$109,762
12	Huntington Beach	City	189744	$42,127	$80,901	$99,038
30	Rancho Santa Margarita	City	47769	$41,787	$104,167	$116,540
25	Mission Viejo	City	93076	$41,436	$96,420	$109,693
20	Lake Forest	City	77111	$39,844	$94,632	$108,211
33	San Juan Capistrano	City	34455	$39,097	$73,806	$86,744
23	Los Alamitos	City	11442	$38,527	$79,861	$90,409
3	Brea	City	38837	$36,195	$81,278	$98,159
18	Laguna Woods	City	16276	$36,017	$35,393	$50,332
9	Fountain Valley	City	55209	$35,487	$81,661	$91,003
21	La Palma	City	15536	$34,475	$84,693	$92,757
5	Costa Mesa	City	109796	$33,800	$65,471	$74,201
38	Tustin	City	74625	$32,854	$73,231	$80,963
7	Cypress	City	47610	$32,815	$82,954	$92,276
28	Orange	City	135582	$32,797	$78,654	$88,423
10	Fullerton	City	134079	$30,967	$69,432	$78,812
29	Placentia	City	50089	$30,451	$78,364	$90,372
19	La Habra	City	60117	$24,589	$63,356	$69,028

4.b) Determine the top 20 cities by per capita income for both counties¶

City	PCI
Laguna Beach	$81,591
Newport Beach	$80,872
Villa Park	$71,697
Laguna Niguel	$51,491
Yorba Linda	$49,485
San Clemente	$47,894
Laguna Hills	$44,751
Seal Beach	$44,115
Irvine	$43,102
Huntington Beach	$42,127
Rancho Santa Margarita	$41,787
Mission Viejo	$41,436
San Juan Capistrano	$39,097
Los Alamitos	$38,527
Camarillo	$37,840
Ojai	$36,769
Moorpark	$36,375
Fountain Valley	$35,487
Simi Valley	$35,467
La Palma	$34,475

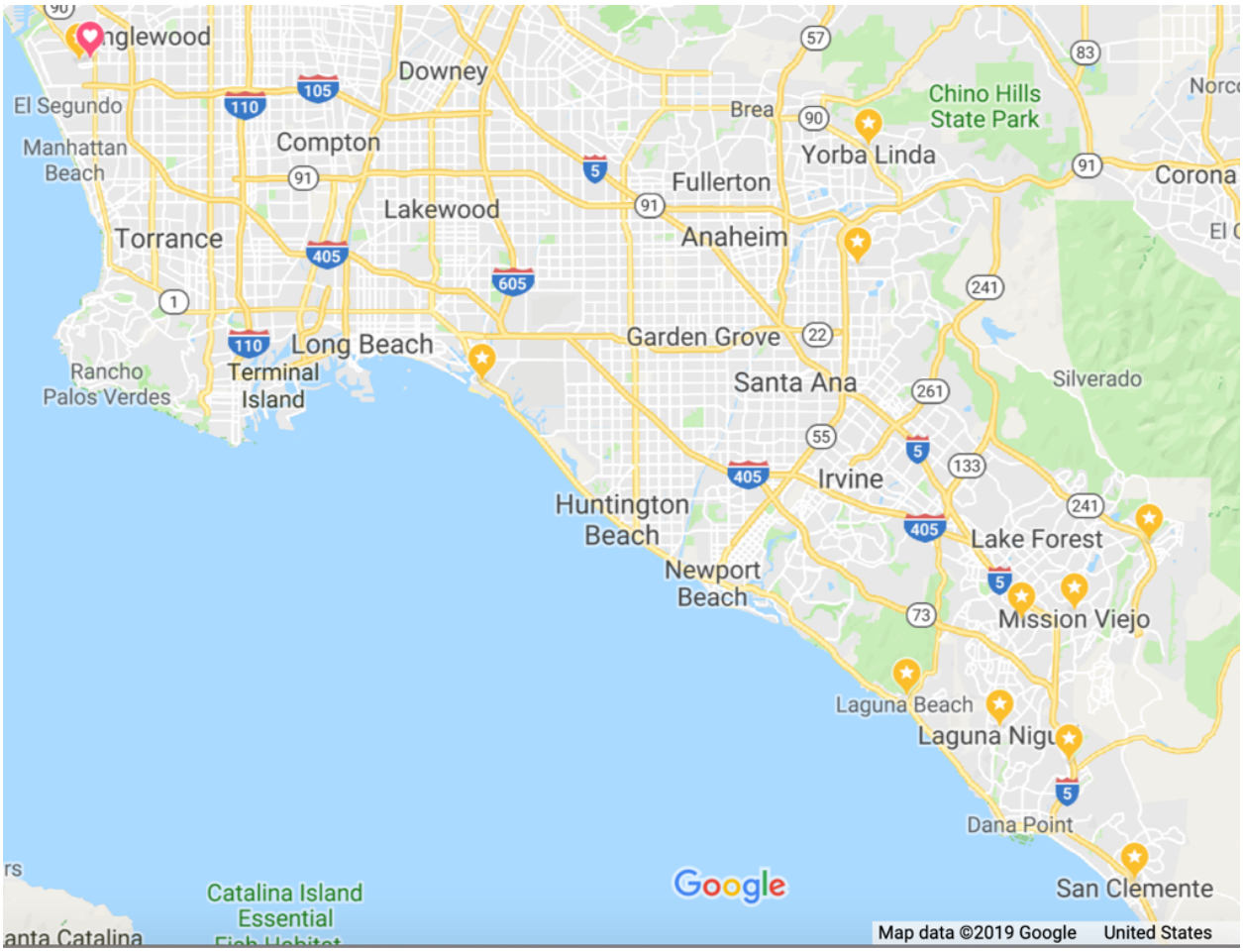
5.a) Determine top 10 cities by crime rate and per capita income in both counties¶

Removed Cities above violent and property crime thresholds

Out[9]:

City	Population	Violent CR	Property CR	PCI
Laguna Beach	23283	2.45	23.54	$81,591
Villa Park	5956	0.34	14.61	$71,697
Laguna Niguel	64533	0.73	11.84	$51,491
Yorba Linda	65820	0.81	11.96	$49,485
San Clemente	65089	1.15	12.89	$47,894
Laguna Hills	31090	0.93	19.94	$44,751
Seal Beach	24764	0.69	22.01	$44,115
Rancho Santa Margarita	49038	0.55	6.51	$41,787
Mission Viejo	95599	0.76	12.52	$41,436
San Juan Capistrano	35449	1.66	14.64	$39,097

5.b) Map the crime rate and per capita income of top 10 cities



Based on selection criteria (crime rate and per capita income) the top 10 cities are all located in Orange county. The heart shape shows the location of Los Angeles International Airport. The Notebook on Github also has several interactive Folium maps.

6. Combine coordinates with crime rate and PCI for chosen cities in California¶

After combining per capita income and crime rates, we narrowed the list of cities to 10.

Sample of cities listed

City	Latitude	Longitude	Zip	Population	Violent CR	Property CR	PCI
Villa Park	33.812662	-117.8162	92861	5956.0	0.34	14.61	$71,697
Placentia	33.640302	-117.7694	92871	51778.0	2.07	17.50	$30,451
Yorba Linda	33.888062	-117.8040	92886	65820.0	0.81	11.96	$49,485
Mission Viejo	33.615462	-117.6409	92692	95599.0	0.76	12.52	$41,436
Tustin	33.741651	-117.8212	92780	77400.0	1.47	21.36	$32,854
.....							

7. Foursquare analysis of city venues¶

Applied Foursquare venue count to top 10 cities.

City	# Venues
Laguna Beach	24
Laguna Hills	6
Laguna Niguel	5
Mission Viejo	1
Rancho Santa Margarita	11
San Clemente	38
San Juan Capistrano	60
Seal Beach	17
Villa Park	6
Yorba Linda	4

8.a) Determine list of finalists based on number of venues¶

Top 5 finalists are:
San Juan Capistrano - 60 venues
San Clemente - 36 venues
Laguna Beach - 24 venues
Seal Beach - 17 venues
Rancho Santa Margarita - 11 venues

8.b) Narrow search by looking at types of venues¶

Top 2 favorite cities
1- Laguna Beach has a park and a beach as venues listed. It is surrounded by a wilderness coastal park.
2- San Juan Capistrano has a historic site and winery. It is further from the beach but is located sufficiently close to the beach and wilderness areas

Cities dropped from the list
Rancho San Margarita has the lowest count of venues and has no parks as venues listed. Among the top 5, it is the furthest from the beach
Seal Beach has no parks as venues listed. Apart from being close to the ocean. It has no redeeming qualities
San Clemente is in 2nd place for number of venues, but it has no parks as venues listed. It is the furthest from the LAX airport among all the cities.
It is located near the beach and wilderness areas. Unfortunately, the city's proximity to the military base makes it less suitable for locating a quiet retreat free of military jets flying overhead.

9. Analyze real estate of two finalists - Laguna Beach and San Juan Capistrano¶

Issues with real estate data

MLS listings are not available as a csv download from web sites unless you work in real estate and have a brokers' license. Web scrapping is not permissible or the data is very transitory and the active link available today will no longer work the following day. You can download historical Zillow real estate data from the data world web site. Redfin offers downloadable market data, but not MLS listings for download. Consequently, it is best to work with what is available namely, market data and apply forecasting methods to this data. You have to capture a snapshot of the data at a given point in time. Then you upload it to an SQL database for further analysis, because the data might not be available the following day.

You can determine the current trends based on historical data from Redfin. Then you forecast outwards from the latest data point in 2019. It is best to look at the MLS data at the very end of the B&B analysis, because of the transitory nature of the data from properties being sold and quickly disappearing from the dataset.

Both SQL datasets have 85 rows for easier comparison. The data was filtered using SQL statements on DB2 to remove years prior to 2012 and reduce file size. Further data wrangling will be done in python after reading the data dictionary and later identifying the columns that contain the most valuable data.

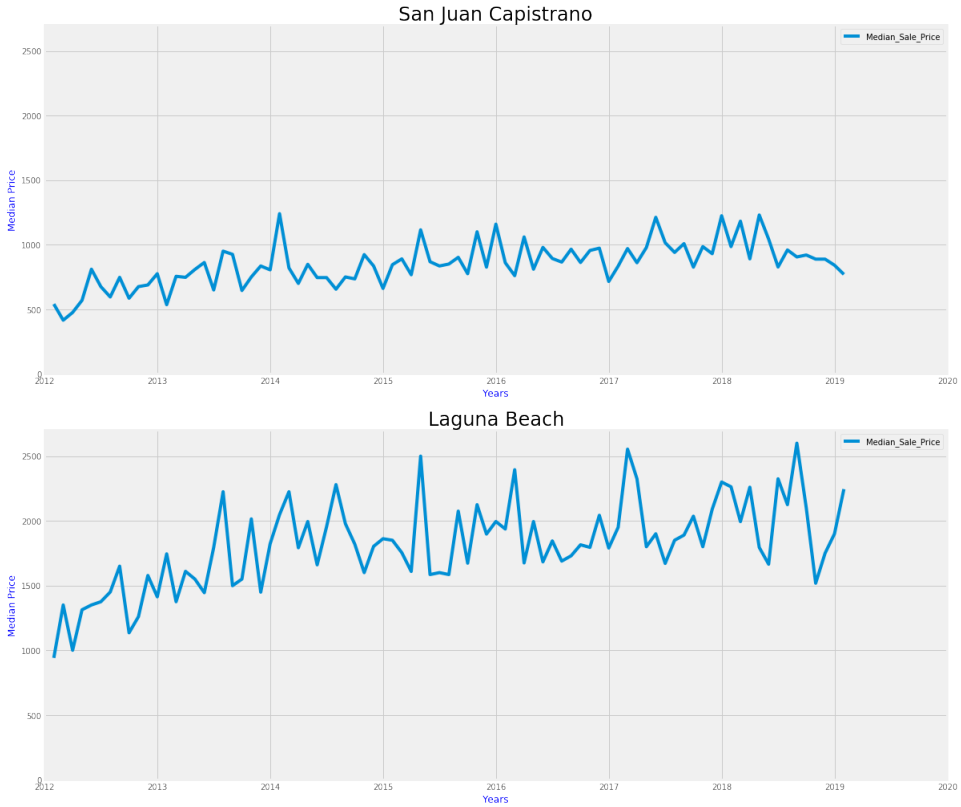
10. Prep datasets for Time Series Analysis¶

Several packages were used to conduct statistical analysis and plotting.
For example,

Matplotlib version: 2.1.0
'ggplot', 'seaborn-white', 'seaborn-ticks', 'classic', 'bmh', 'grayscale', '_classic_test', 'seaborn-notebook', 'seaborn', 'fivethirtyeight', 'seaborn-bright', 'seaborn-dark', 'seaborn-poster', 'seaborn-deep', 'seaborn-whitegrid', 'seaborn-talk', 'Solarize_Light2', 'fast', 'seaborn-muted', 'seaborn-colorblind', 'seaborn-dark-palette', 'seaborn-paper', 'dark_background', 'seaborn-darkgrid', 'seaborn-pastel'

11. Look at historical data of median sale prices¶

11.a) Compare Laguna Beach vs. San Juan Capistrano median sale prices¶



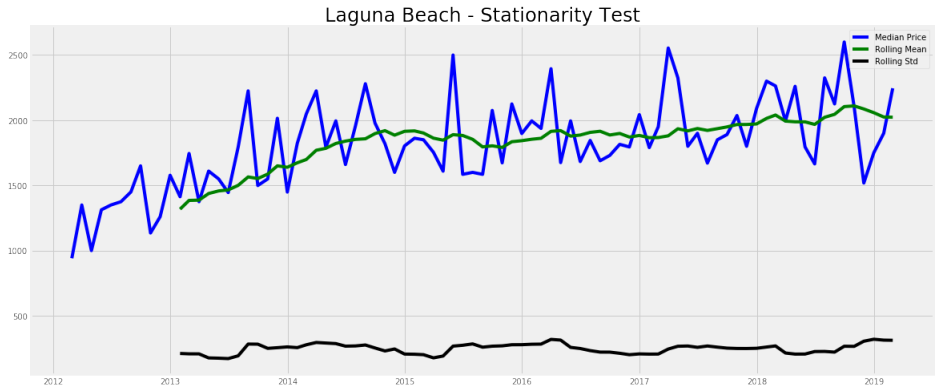
Overall, the median sale prices of Laguna Beach are much higher than San Juan Capistrano. This is due to Laguna Beach's closer proximity to the beach than San Juan Capistrano. The prices fluctuations are greater for Laguna Beach than San Juan Capistrano. Finally, we see that San Juan Capistrano's median sale prices have declined whereas Laguna Beach median prices have rebounded in late December 2018. This favors San Juan Capistrano as a potential location for buying opportunities. Further analysis is required to identify any presence of seasonality or a trend from the data.

11.b) Prepare data for stationarity evaluation of datasets¶

Before applying a time series model, we must show that it is stationary. For example, the statistical properties (mean and variance) should remain constant over time. Looking at the data's behavior over time, it must be the same in the future to accurately forecast the series. The statistical properties can be a constant mean, a constant variance or an auto co-variance that do not depend on time. The most common method for testing stationarity is to use the Dickey-Fuller test. First, we determine a null hypothesis, namely that the time series is non-stationary. Second, we evaluate the different statistical confidence levels. The test results should indicate that the test statistics are less than the critical values. If true, we can reject the null hypothesis and say that the time series is indeed stationary. In the next few chapters of this project, we'll observe different techniques to improve data quality in preparation for time series

analysis and forecasting. However, please keep in mind that applying all of the techniques will not necessarily provide the best results. There is a higher probability of over-fitting where too much noise is removed from the data and the forecast is no longer useful. It is best to test incrementally each transformation to validate stationarity before applying the next data transformation.

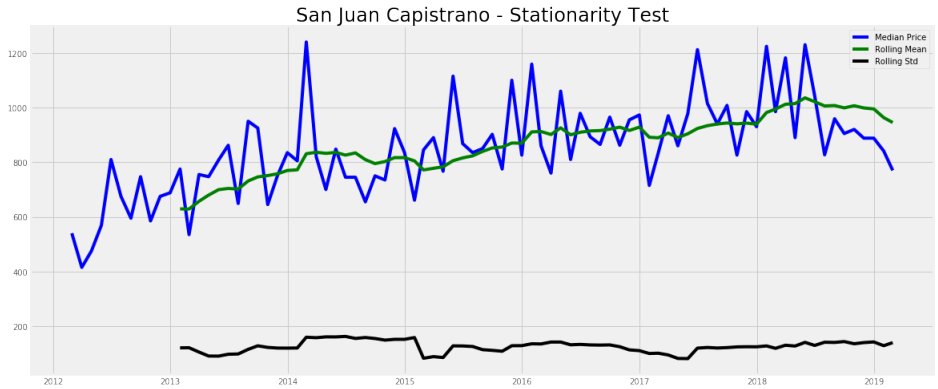
11.c) Display results of Dickey-Fuller Tests¶



Test Results:

Test Stat -5.909377e+00
p-value 2.661176e-07
#Lags Used 0.000000e+00
Number of Observations 8.400000e+01
Critical Value (10%) -2.585482e+00
dtype: float64
Test Stat -5.909377e+00
p-value 2.661176e-07
#Lags Used 0.000000e+00
Number of Observations 8.400000e+01
Critical Value (10%) -2.585482e+00
Critical Value (1%) -3.510712e+00
dtype: float64
Test Stat -5.909377e+00
p-value 2.661176e-07
#Lags Used 0.000000e+00
Number of Observations 8.400000e+01
Critical Value (10%) -2.585482e+00
Critical Value (1%) -3.510712e+00
Critical Value (5%) -2.896616e+00
dtype: float64

Looking at the stationarity test results we see that the Laguna Beach dataset is stationary, because the Test Stat value of -5.909377e+00 is less than the Critical Values (1%, 5% and 10%).



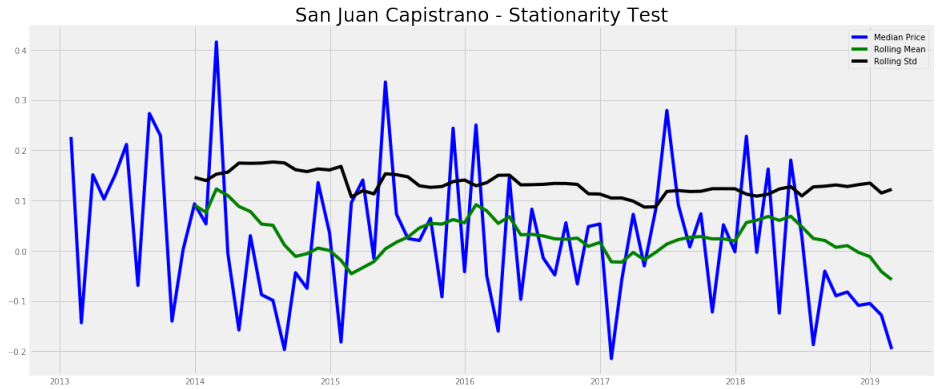
Test Results:

Test Stat -3.403302
p-value 0.010847
#Lags Used 2.000000
Number of Observations 82.000000
Critical Value (10%) -2.585949
dtype: float64
Test Stat -3.403302
p-value 0.010847
#Lags Used 2.000000
Number of Observations 82.000000
Critical Value (10%) -2.585949
Critical Value (1%) -3.512738
dtype: float64
Test Stat -3.403302
p-value 0.010847
#Lags Used 2.000000
Number of Observations 82.000000
Critical Value (10%) -2.585949
Critical Value (1%) -3.512738
Critical Value (5%) -2.897490
dtype: float64

Laguna Beach dataset is stationary, but the San Juan Capistrano dataset is not. The Test Stat is less than the Critical Value at (5% and 10%) but not at the 1% value level. Therefore, we need to make the San Juan Capistrano dataset stationary. It is likely that the dataset has seasonality or a trend indicated by a non-constant mean.

11.d) Smooth & dampen trend noise of San Juan Capistrano using logarithmic transformation & moving average¶

We can use one of three transformations to reduce the effects of the trend on stationarity, namely logarithmic, square-root or cubic root. These transformations will reduce the largest values relative to the smaller ones. This analysis will use the logarithmic transformation. To remove unwanted noise from the dataset, we apply a rolling or moving average. After applying these transformations, we test the dataset again and look for improvements in stationarity.

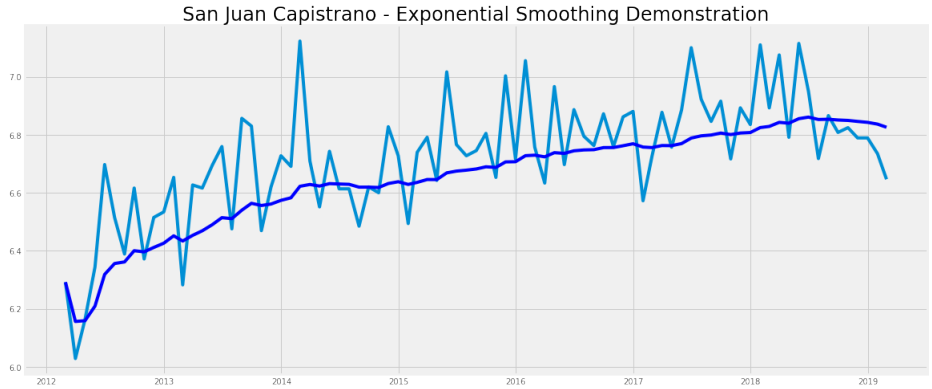


Test Results:

Test Stat -8.509962e+00
p-value 1.174726e-13
#Lags Used 0.000000e+00
Number of Observations 7.300000e+01
Critical Value (10%) -2.588371e+00
dtype: float64
Test Stat -8.509962e+00
p-value 1.174726e-13
#Lags Used 0.000000e+00
Number of Observations 7.300000e+01
Critical Value (10%) -2.588371e+00
Critical Value (1%) -3.523284e+00
dtype: float64
Test Stat -8.509962e+00
p-value 1.174726e-13
#Lags Used 0.000000e+00
Number of Observations 7.300000e+01
Critical Value (10%) -2.588371e+00
Critical Value (1%) -3.523284e+00
Critical Value (5%) -2.902031e+00
dtype: float64

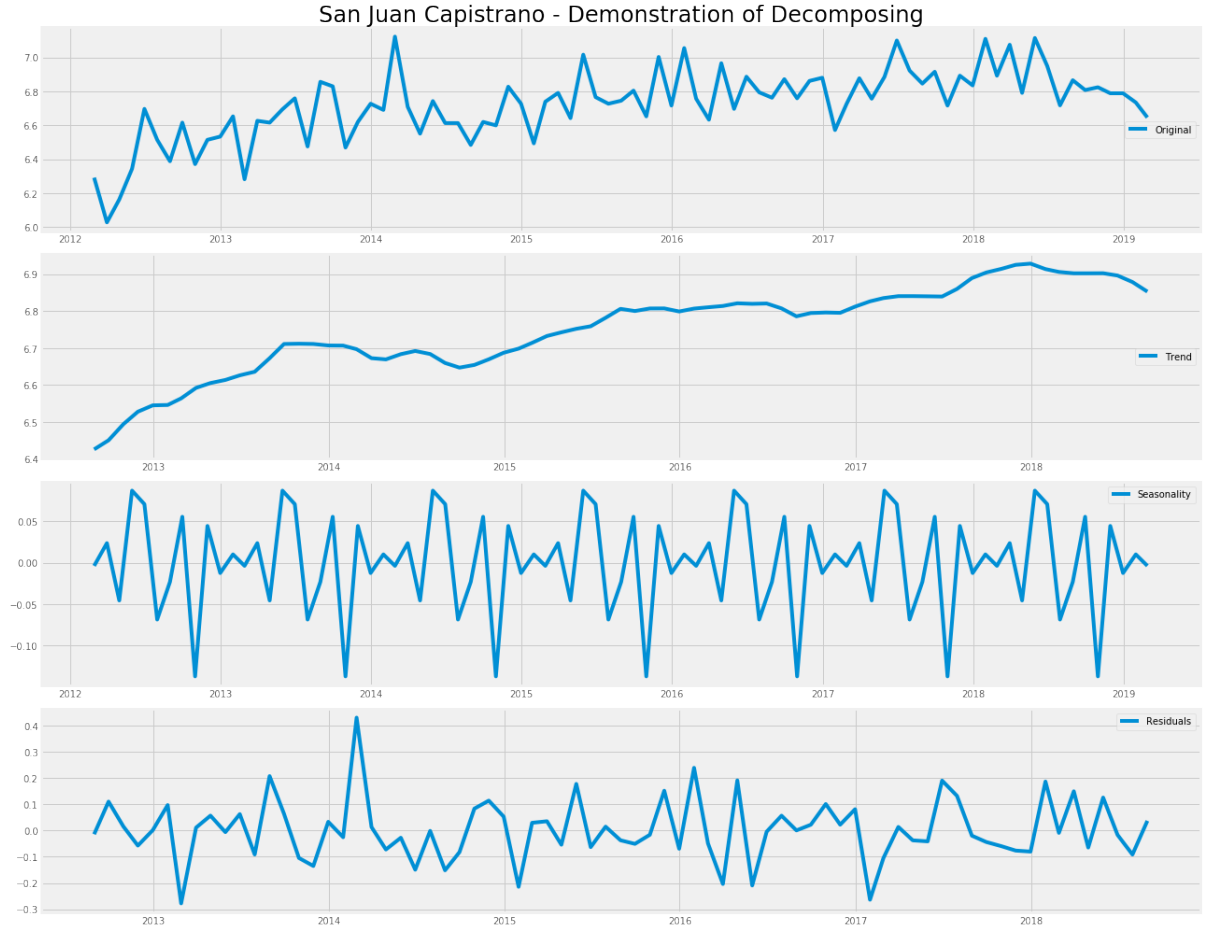
Looking at the test stat, we now see that San Juan Capistrano dataset is stationary for all value levels. Therefore, we were successful at reducing the noise in the data and seeing stationarity. The Test Stat is now lower than the Critical Value at all levels including the 1% critical value. Looking at the rolling mean, we do not see a trend.

11.e) Find trend of San Juan Capistrano dataset by applying exponential smoothing¶

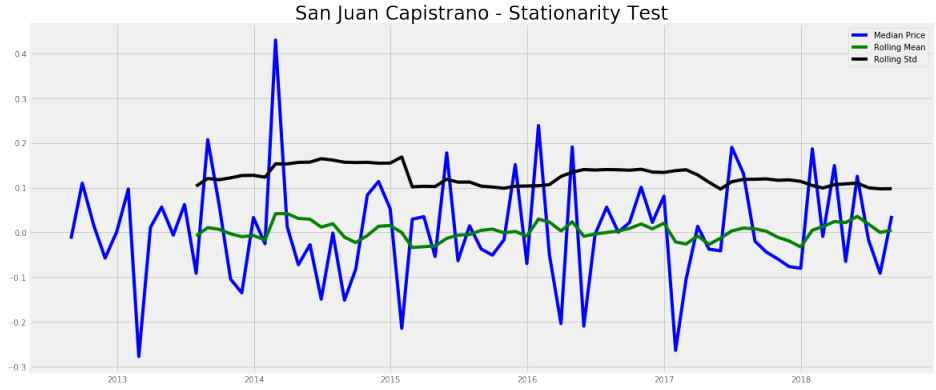


The test results have already shown that the dataset for San Juan Capistrano has stationarity. For demonstration purposes, if we needed to further transform the data, we could apply exponential smoothing as shown in the chart above. Looking at the chart, we now see that San Juan Capistrano dataset shows an upward a trend.

11.f) Remove trend and seasonality of dataset by applying differencing and decompose¶



The test results have already shown that the dataset for San Juan Capistrano has stationarity. For demonstration purposes, if we needed to further transform the data, we could apply decomposing as shown in the four charts above. We would then use the residuals data to test for stationarity.



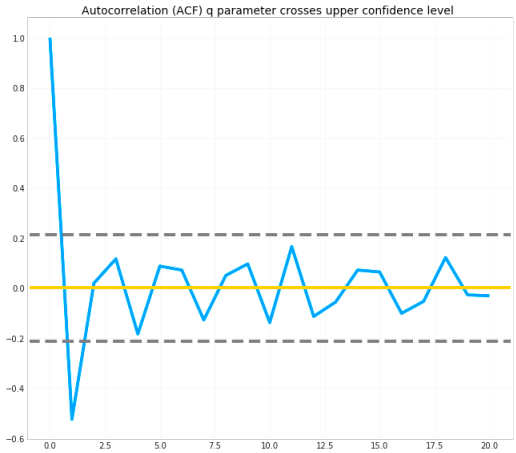
Test Results:

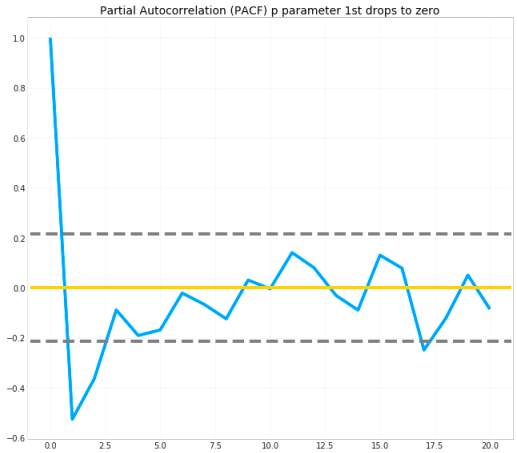
Test Stat -5.799221e+00
p-value 4.673420e-07
#Lags Used 6.000000e+00
Number of Observations 6.600000e+01
Critical Value (10%) -2.590724e+00
Test Stat -5.799221e+00
p-value 4.673420e-07
#Lags Used 6.000000e+00
Number of Observations 6.600000e+01
Critical Value (10%) -2.590724e+00
Critical Value (1%) -3.533560e+00
dtype: float64
Test Stat -5.799221e+00
p-value 4.673420e-07
#Lags Used 6.000000e+00
Number of Observations 6.600000e+01
Critical Value (10%) -2.590724e+00
Critical Value (1%) -3.533560e+00
Critical Value (5%) -2.906444e+00
dtype: float64

The test results have already shown that the dataset for San Juan Capistrano has stationarity. Looking at the Test Stat, we now see that San Juan Capistrano dataset is stationary for all value levels. Therefore, we were successful at reducing the noise in the data and seeing stationarity. The Test Stat is now lower than the Critical Value at all levels including the 1% Critical Value. Looking at the rolling mean, we do not see a trend.

11.g) Plot confidence intervals using autocorrelation and partial correlation for setting p & q values¶

Plotting the confidence intervals allows use to determine the optimal parameters for forecasting a time series.





Looking at the autocorrelation charts above, we see the values for both parameters P and Q should be no more than 2 in an Auto Regressive Integrated Moving Average (ARIMA) model. In short, the parameters (p, d, q) help determine the predictors of the linear regression equation.

The two gray dotted lines on either sides of 0 shown in yellow are the confidence intervals. We use these to determine the 'p' and 'q' values as q indicates first time where the PACF crosses the upper confidence interval. Looking at these charts we observe that the interval is close to 1. Therefore, p = 1 and the parameter q being the first time where the ACF crosses the upper confidence interval. In this case, it is near 1 and q becomes 1.

P is the number of AR (Auto-Regressive) terms. P = Periods to lag. If P= 3 then we will use the three previous periods of our time series in the autoregressive portion of the calculation). P helps adjust the line that is being fitted to forecast the series. As an equation, if p is 3 then the predictor for y(t) will be y(t-1),y(t-2),y(t-3).

Q is a variable that denotes a lag of the error component, where error component is a part of the time series undefined by trend or seasonality. The parameter q is the number of MA (Moving-Average) terms. In an equation,if q is 3 then the predictor for y(t) will be y(t-1),y(t-2),y(t-3).

D is the transformation of a time series into a stationary one(series without trend or seasonality) by using the differencing method. D refers to the number of differencing transformations required by the time series to become stationary.

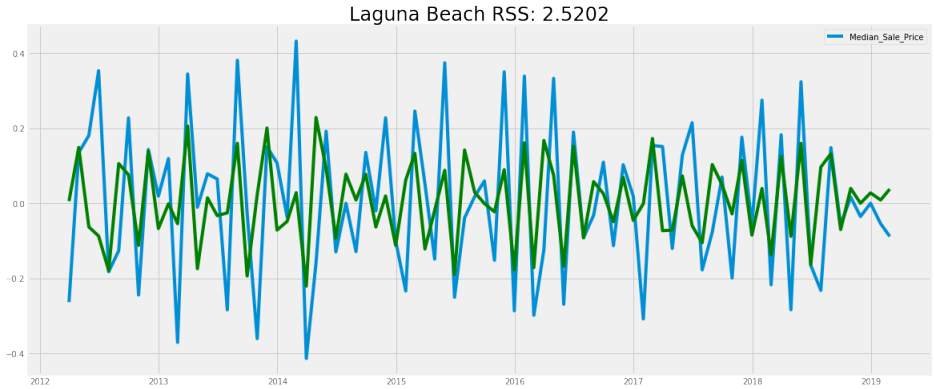
In regard to the Autocorrelation function (ACF), the x axis of the ACF plot indicates the lag at which the autocorrelation is calculated; the y axis displays the value of the correlation (between −1 and 1). A partial correlation (PACF) is defined as a conditional correlation. It is the correlation between two variables under the assumption that we also consider the values of some other set of variables. The partial autocorrelation function (PACF) provides the partial correlation of a stationary time series with its own lagged values. This contrasts with the autocorrelation function, which ignores other lags.

12. Apply forecasting models AR, MA and ARIMA¶

Autocorrelation refers to how correlated a time series is with the past values. We use an ACF plot to view the correlation between the points, including the lag unit. In reference to the ACF model, the correlation coefficient is in the x-axis whereas the number of lags is shown in the y-axis. Usually with an ARIMA model, we make use of either the AR model or the MA model results. Rarely, do we use the results from both. We use the ACF plot to decide which one of the parameters is best suited for our time series model.

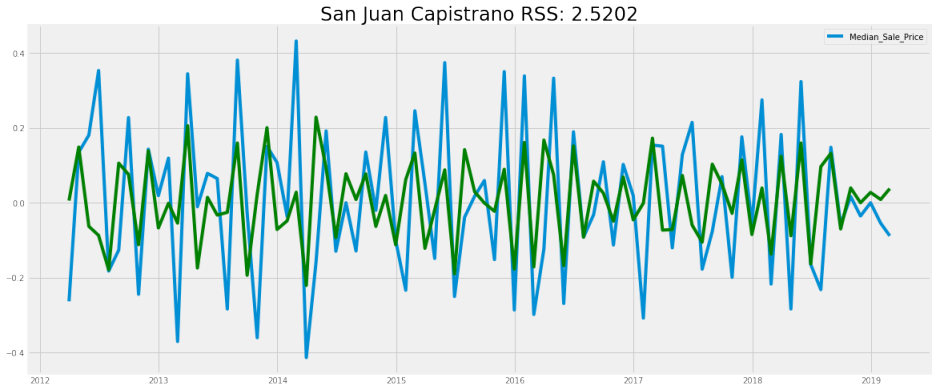
12.a) Test AR Model on Laguna Beach dataset¶

Test Results: (0.5,1,'Laguna Beach RSS: 2.5202')



12.b) Test AR Model on San Juan Capistrano dataset¶

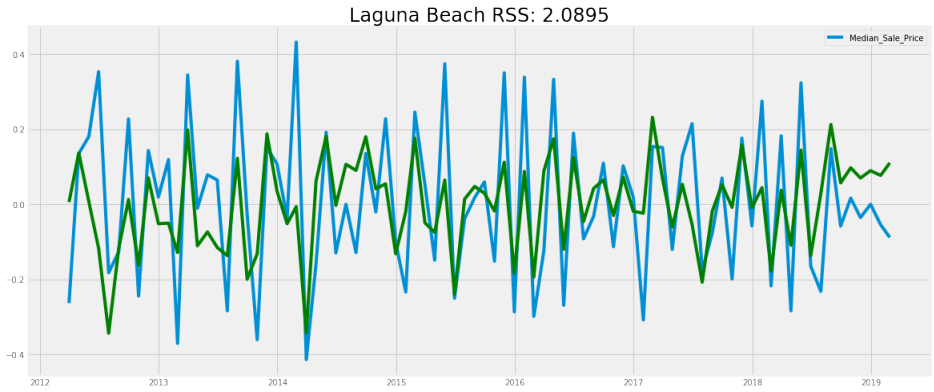
Test Results: (0.5,1,'San Juan Capistrano RSS: 2.5202')



The residual sum of squares (RSS) is used to help you determine if your statistical model is a good fit for the data. It measures the difference between your data and the predicted values in your estimation model (a "residual" is a measure of distance from a given data point to the regression line). Looking at the first set of charts using the AR model, we notice that the residual sum of squares (RSS) values is 2.52 for both datasets. The lower the value of RSS, the better the fit. Hopefully we can achieve a tighter fitting model using a different forecasting model, like MA or ARIMA.

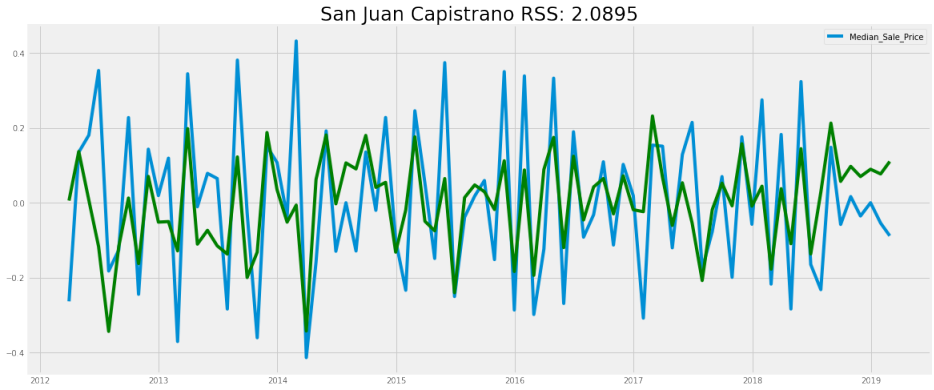
12.c) Test MA Model on Laguna Beach dataset¶

Test Results: (0.5,1,'Laguna Beach RSS: 2.0895')



12.d) Test MA Model on San Juan Capistrano dataset¶

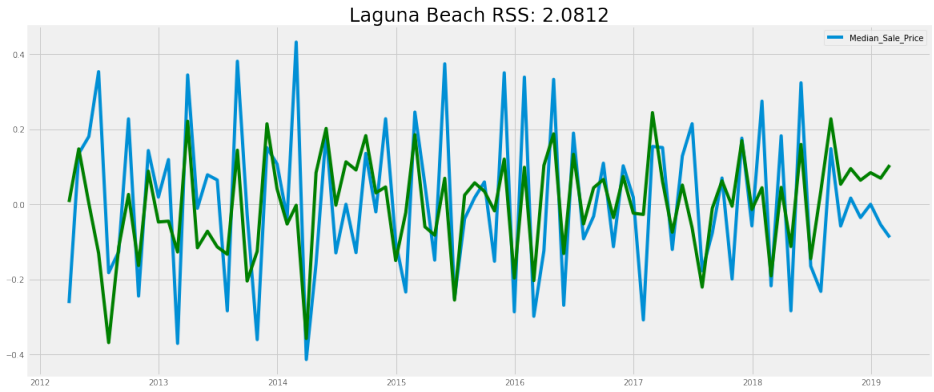
Test results: (0.5,1,'San Juan Capistrano RSS: 2.0895')



Looking at the second set of charts using the MA model, we notice that the residual sum of squares (RSS) values denotes an improvement with a lower RRS at 2.08 for both datasets. Hopefully we can achieve a tighter fitting model using the ARIMA model.

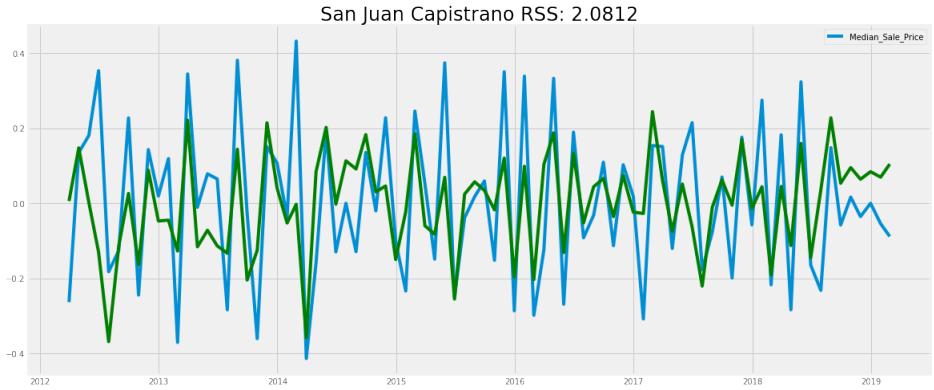
12.e) Test ARIMA Model on Laguna Beach dataset¶

Test Results: (0.5,1,'Laguna Beach RSS: 2.0812')



12.f) Test ARIMA Model on San Juan Capistrano dataset¶

Test Results: (0.5,1,'San Juan Capistrano RSS: 2.0812')



After evaluating all the ARIMA models, we notice a marginal improvement of the RSS values for both datasets.

13. Use SARIMAX forecasting method¶

The SARIMA model is similar to the ARIMA model. However, we add in a few parameters to account for the seasons

ARIMA (p, d, q)(P, D, Q) m,

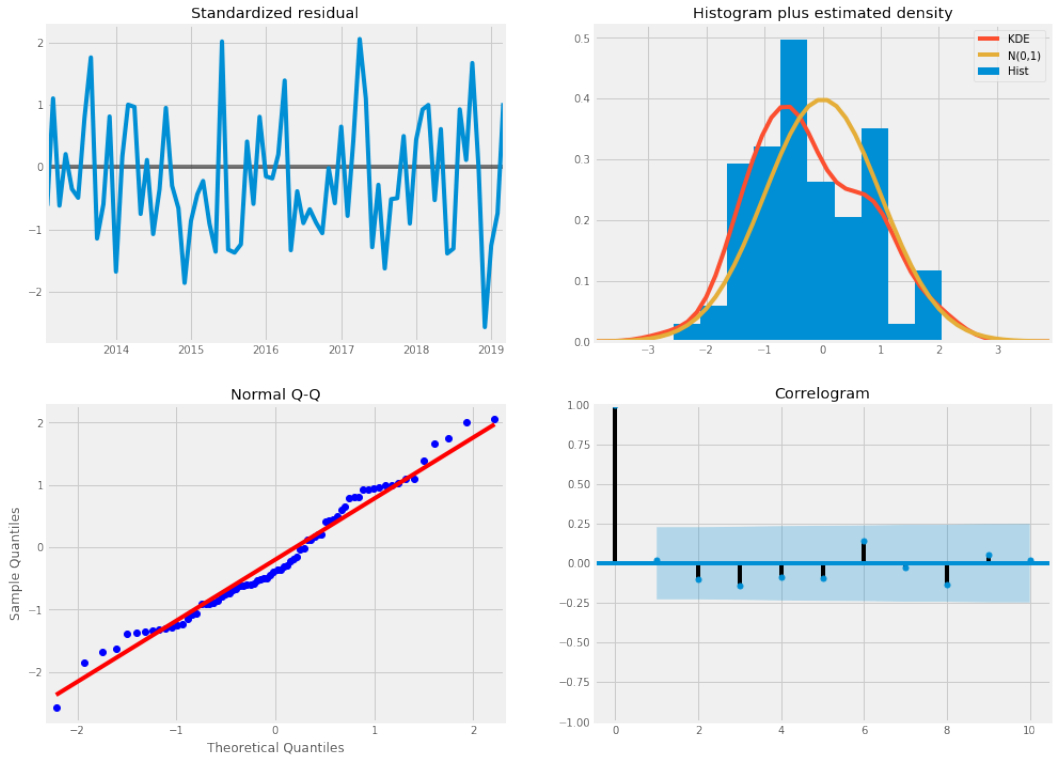
p—is the order (number of time lags
d—is the degree of differencing
q—is the order of moving average terms
m—refers to the number of periods in each season
(P, D, Q)— represents the (p, d, q) for the seasonal part of the time series

13.a) Select p. d. q. parameters for time series model for Laguna Beach¶

13.b) Assess soundness of time series model¶

```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8498      0.081    -10.509      0.0      -1.008      -0.691
ma.S.L4       -1.0002      0.215     -4.659      0.0      -1.421      -0.579
sigma2         7.3e+04   2.9e-06   2.5e+10      0.0    7.3e+04    7.35e+04
==============================================================================
```

The table above shows a key column, namely the *coef* column. It shows the importance of each feature like standard deviation and z score and how each one impacts the time series. The P>|z| column tells us of the relative weighting of each feature. We see that each weight has a p-value lower to 0.05, so we can include them in our forecasting model.



The primary goal is of this model is to ensure that it has uncorrelated and normally distributed with zero-mean residual values. Looking at the charts above, it suggests that the model residuals are normally distributed.

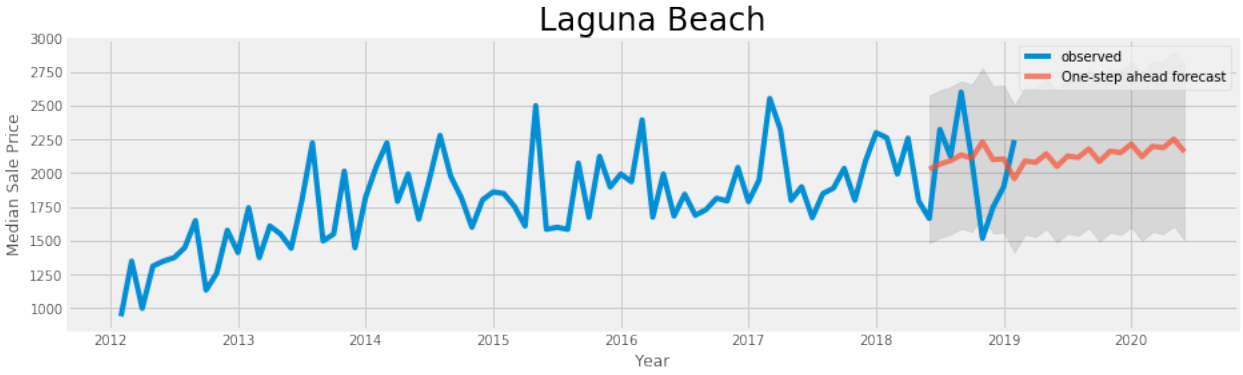
Evidence of this appears in the top right plot, we see the red KDE line closely tracks the path of the N(0,1) line. This line is the standard notation for a normal distribution curve showing a mean of 0 and a standard deviation of 1. It clearly indicates a normal distribution of residuals. Evidence of nnother strong indication of this is seen by looking at the qq-plot on the bottom left, we see an ordered distribution of residuals as blue dots. They follow the samples linear trend taken from

a normal distribution with N(0,1). The residuals shown on the top left plot do not display seasonality and are likely white noise. Finally, the autocorrelation shown on the correlogram plot on the bottom right, show evidence that the residuals have a low correlation with lagged versions. In summary, all these observations point to a good working model for forecasting future median sales values for Laguna Beach.

13.c) Create a table of values for time series model¶

Period Ending	lower Median_Sale_Price	upper Median_Sale_Price
2020-02-29	1501.038725	2744.739124
2020-03-31	1566.686656	2832.390584
2020-04-30	1549.316051	2828.306634
2020-05-31	1606.315419	2898.511009
2020-06-30	1506.358949	2811.740861

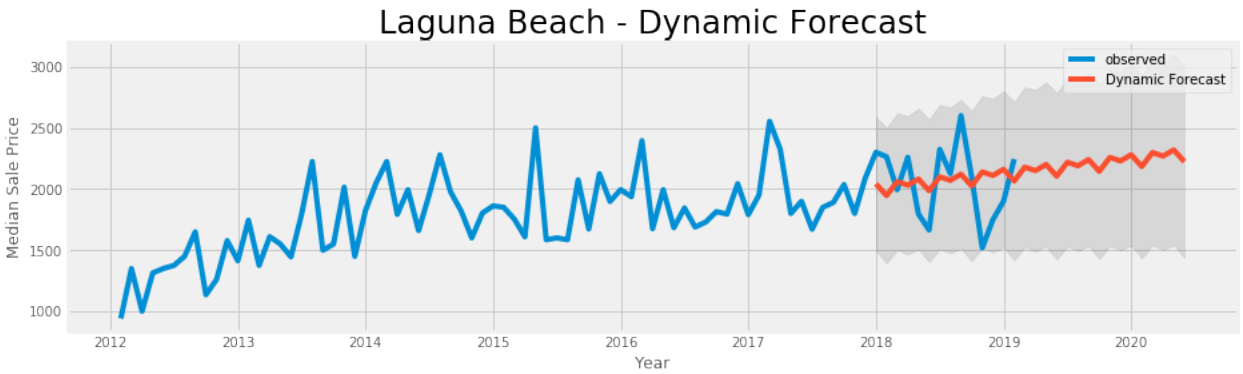
13.d) Plot the forecasted values with historical data¶



13.e) Plot dynamic forecast for Laguna Beach¶

Mean Squared Error of forecast : 1415523.59

Period Ending	lower Median_Sale_Price	upper Median_Sale_Price
2018-01-31	1495.127530	2589.715551
2018-02-28	1392.648767	2499.559123
2018-03-31	1500.355082	2619.768050
2018-04-30	1465.401142	2596.716888
2018-05-31	1505.263044	2658.874472



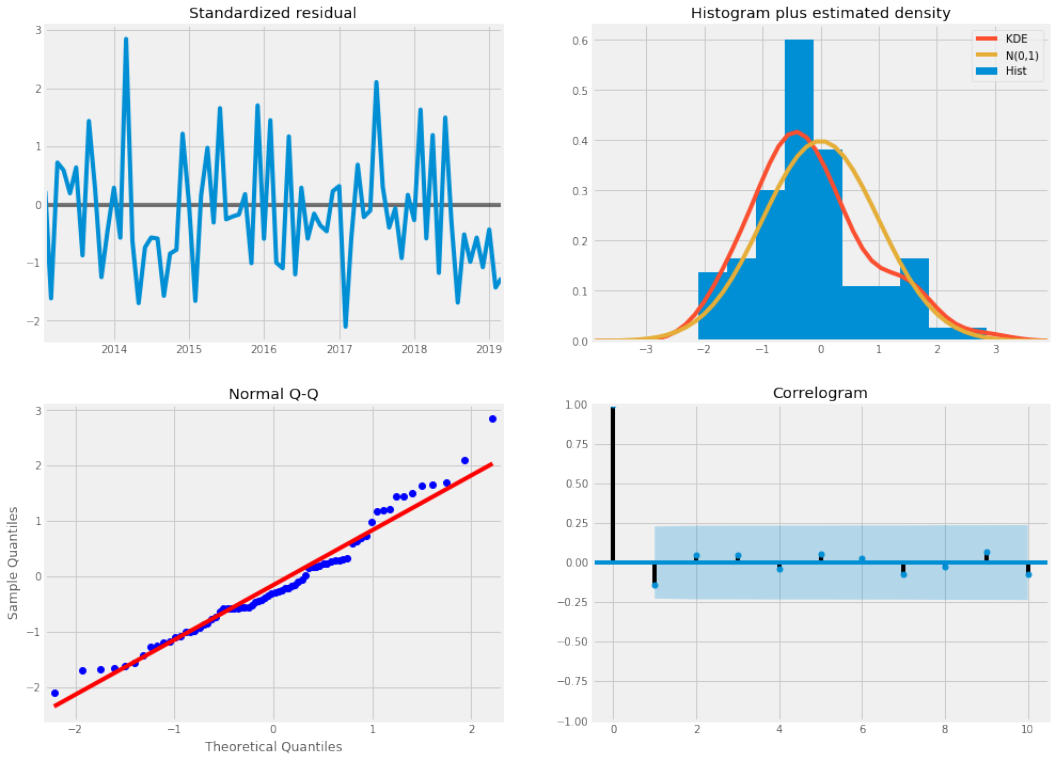
After performing forecasting on the median sale prices of Laguna Beach we notice a healthy real estate market with prices continuing to climb in 2020. The close proximity to the Pacific ocean makes this location relatively expensive to other cities nearby. The median sale price is expected to reach approximately $2.3 million in 2020.

13.f) Select p.d.q parameters for time series model for San Juan Capistrano¶

13.g) Assess soundness of time series model¶

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.7872	0.106	-7.398	0.000	-0.996	-0.579
ar.S.L4	-0.1998	0.134	-1.488	0.137	-0.463	0.063
ma.S.L4	-1.0000	0.178	-5.631	0.000	-1.348	-0.652
sigma2	1.698e+04	1.05e-05	1.6e+09	0.000	1.7e+04	1.7e+04

The table above shows a key column, namely the *coef* column. It shows the importance of each feature like standard deviation and z score and how each one impacts the time series. The P>|z| column tells us of the relative weighting of each feature. We see that each weight has a p-value lower to 0.05, so we can include them in our forecasting model.



The primary goal is of this model is to ensure that it has uncorrelated and normally distributed with zero-mean residual values. Looking at the charts above, it suggests that the model residuals are normally distributed.

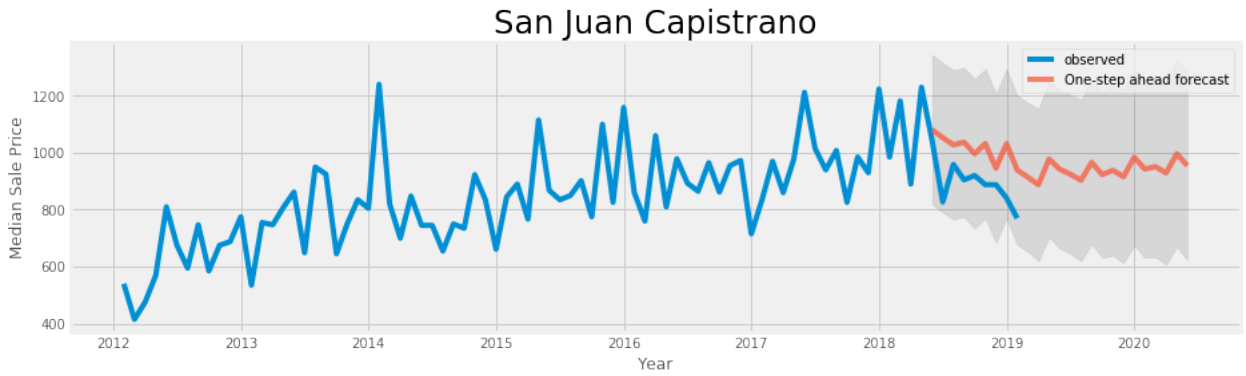
Evidence of this appears in the top right plot, we see the red KDE line closely tracks the path of the N (0,1) line. This line is the standard notation for a normal distribution curve showing a mean of 0 and a standard deviation of 1. It clearly indicates a normal distribution of residuals. Another strong indication of this is seen by looking at the "qq-plot" on the bottom left, we see an ordered distribution of residuals as blue dots. They follow the samples linear trend taken from a normal distribution with N (0,1). The residuals shown on the top left plot do not display seasonality and are likely white noise. Finally, the autocorrelation shown on the correlogram plot on the bottom right, indicate that the residuals have a low correlation with lagged versions. In summary, all these observations point to a good working model for forecasting future median sales values for San Juan Capistrano.

13.h) Create a table of values for time series model¶

Out[66]:

Period Ending	lower Median_Sale_Price	upper Median_Sale_Price
2020-02-29	630.828872	1255.102226
2020-03-31	633.497360	1269.037790
2020-04-30	607.004853	1251.257156
2020-05-31	669.642070	1322.629199
2020-06-30	623.813548	1285.658409

13.i) Plot the forecasted values with historical data¶

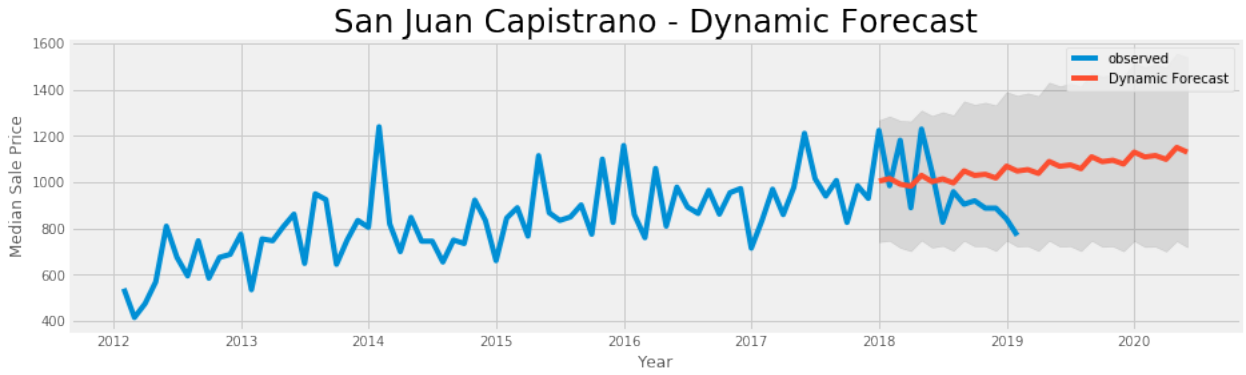


13.j) Plot dynamic forecast for San Juan Capistrano¶

Mean Squared Error of forecast : 28694.93

Out[68]:

Period Ending	lower Median_Sale_Price	upper Median_Sale_Price
2018-01-31	741.814316	1268.227508
2018-02-28	747.056458	1285.434194
2018-03-31	716.784853	1266.523540
2018-04-30	702.529455	1263.531417
2018-05-31	748.861970	1311.006321



After performing forecasting on the median sale prices of San Juan Capistrano we notice a healthy real estate market with prices continuing to climb in 2020. Despite the recent decline, we foresee a high rise in median sale prices. The recent dip provides a buying opportunity. The median sale price is expected to reach approximately $1.1 million in 2020. This is considerably cheaper than a similar property in Laguna Beach. In a nutshell, the real estate market is most favorable in San Juan Capistrano

14. Competitive Analysis¶

14.a) Category Search¶

Doing a Venue Category Id search on Foursquare we get limited results for San Juan Capistrano. we use the following categories with Hotel being the parent of the category hierarchy for accommodations.

Hotel: 4bf58dd8d48988d1fa931735
Bed & Breakfast: 4bf58dd8d48988d1f8931735
Boarding House: 4f4530a74b9074f6e4fb0100
Inn: 5bae9231bedf3950379f89cb
Motel: 4bf58dd8d48988d1fb931735
Resort: 4bf58dd8d48988d12f951735
Vacation Rental: 56aa371be4b08b9a8d5734e1

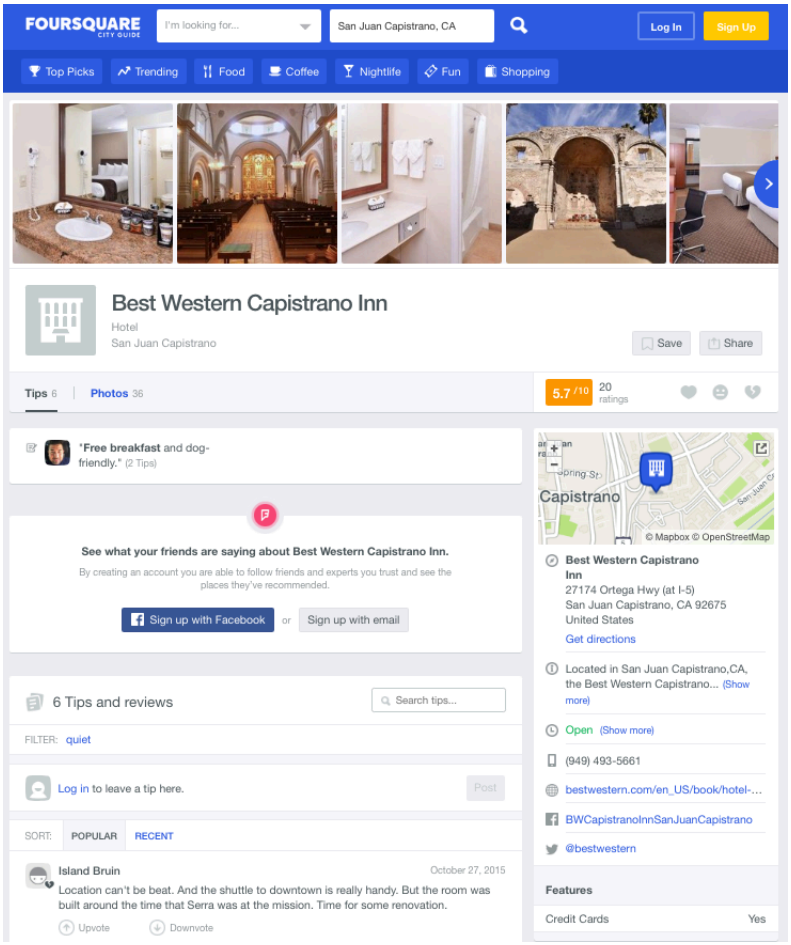
Out[70]:

	name	categories	lat	lng
0	Best Western Capistrano Inn	Hotel	33.502044	-117.656502

Out[71]:

	name	categories	lat	lng
0	Sundried Tomato Cafe	Italian Restaurant	33.501037	-117.662777

14.b) Evaluate Competitors¶



After searching all the Hotel categories, we come up with only one competitor, namely the Best Western Capistrano Inn. Let's size up our competitor by looking at the Foursquare ratings for the Best Western Capistrano Inn.

This establishment has a rating of out of 5.7 / 10 based on 6 responses

5.7. The ratings for this Inn are far from stellar. Let's visit the Best Western Capistrano Inn on Foursquare to get all the details. Like many web sites, Foursquare will prevent you from displaying their web page in an i-frame. Therefore, one is forced to display a screen capture of the web page instead. You will find the link to the page here https://foursquare.com/v/best-western-capistrano-inn/4ba545d9f964a52088f438e3

In summary, the competitive landscape of San Juan Capistrano is very favorable. We should have a high probability of success in beating the competition with our new Bed and Breakfast "Auberge Beau Sejour".

15. Evaluate MLS listings¶

15.a) Build Key Features List to Filter MLS listings¶

Looking at the median prices of home listings they fall within range of the investment budget of 4 million.

Comparing Laguna Beach to San Juan Capistrano, The median price for San Juan Capistrano is trending downwards
whereas Laguna Beach is moving in the opposite direction. Furthermore the median prices in San Juan Capistrano
are lower than those of Laguna Beach. Therefore, San Juan Capistrano is now the preferred location to find a property.

Two properties identified based on the following property search filter criteria:
property-type=house
max-price=4M
min-beds=5
min-baths=5
min-sqft=5k-sqft
min-parking=5
min-lot-size=1-acre
has-pool

15.b) Compare Property Specs of Chosen Properties¶

Based on this criterion, there are two properties on the market in San Juan Capistrano, CA.

Out[76]:

	sale_type	property_type	address	city	zip	price	beds	baths	lot_size	SQFT	mls_
0	MLS Listing	Single Family Residential	27902 Via Madrina	San Juan Capistrano	92675	3498000	5	6.00	64791	417	OC18268633
1	MLS Listing	Single Family Residential	32061 Cook Lane Ln S	San Juan Capistrano	92667	3250000	6	6.25	43560	487	PW18235467

15.c) Review Photos and Details of Properties¶

32061 Cook Lane Ln S

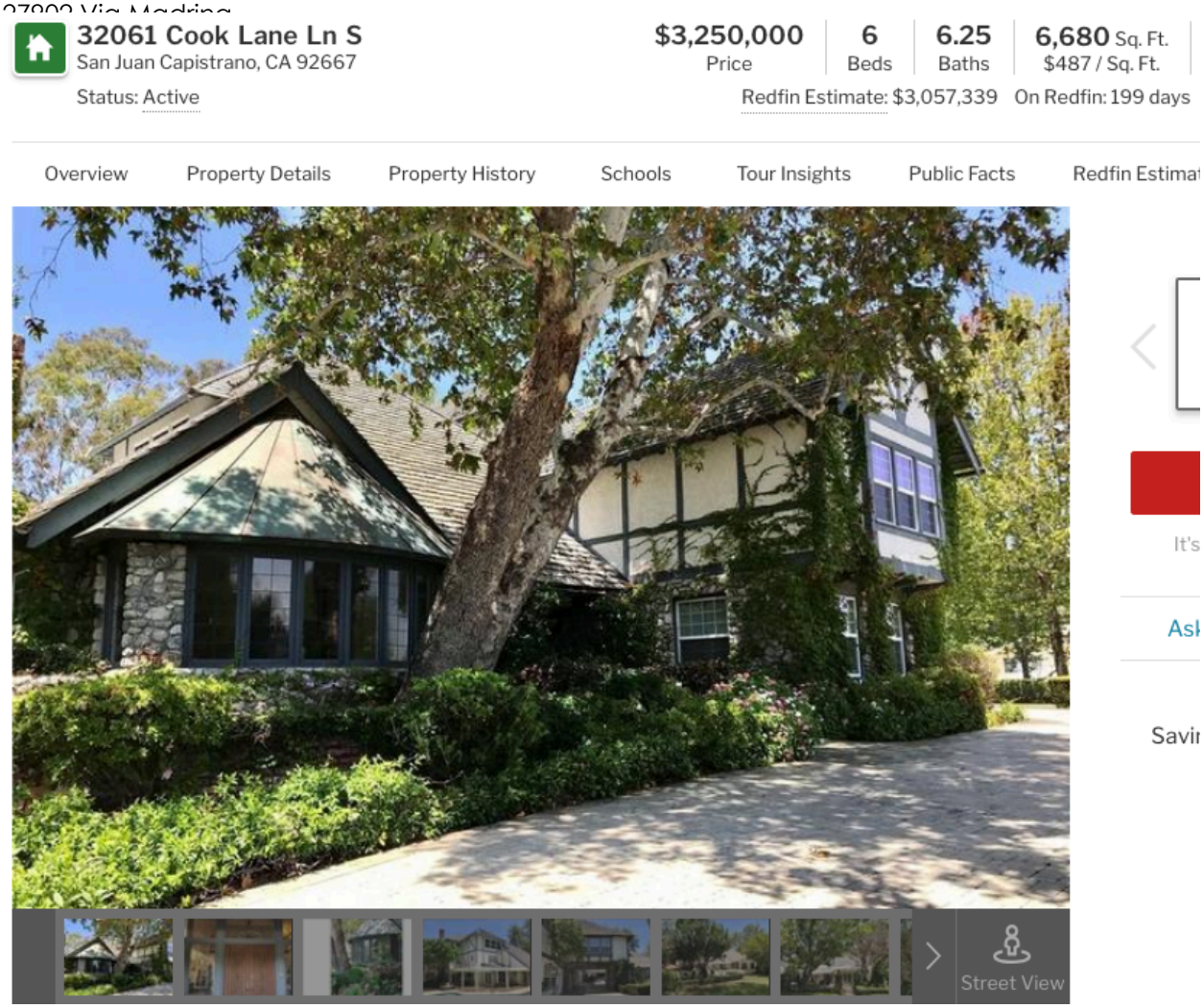
27902 Via Madrina
San Juan Capistrano, CA 92675
Status: Active

$3,498,000	5	6	8,380 Sq. Ft.
Price	Beds	Baths	$417 / Sq. Ft.

Redfin Estimate: $3,258,009 On Redfin: 159 days

Overview | Property Details | Property History | Schools | Tour Insights | Public Facts | Redfin Estima

27902 Via Madrina



Out[77]:

	address	price	beds	baths	lot_size	Sqft
0	27902 Via Madrina	3498000	5	6.00	64791	417
1	32061 Cook Lane Ln S	3250000	6	6.25	43560	487

A quick comparison of these two properties, we see that the property on Cook Ln. appears to be the best choice.

However, property specs are not enough to make a final decision on which one to buy. It is best to see pictures and visit the two potential properties listed on Trulia.

Hopefully, the local zoning laws will allow the conversion of this residential property into a B&B. If there are restrictions, then we can expand the search to new areas within San Juan Capistrano where B&B businesses are permitted.

15.d) Conclusion of Real Estate Market Analysis and Finalists to Establish a B&B¶

We started this analysis by investigating the crime rates across California and came up with four counties.
Then we cut the number of counties from four to two, by looking at driving distance to LAX with a maximum driving time of 2 hours.
Then we reviewed the crime rates of these two counties namely Ventura and Orange.
We narrowed the search to 18 cities by combining the crime rates and the per capita income of each city within these two counties.
Then we looked at the available venues of each of these cities and reduced the list further to 5 cities within Orange county.
Then we reduced the list to two cities after considering access to the beach and closeness to regional parks.
The next stage involved an analysis of the real estate markets for these two finalists.
By comparing the median pricing trends of each city, San Juan Capistrano beat out Laguna Beach.
A competitive analysis of San Juan Capistrano, also provided positive results for the new owner.
Only one poorly rated competitor was located in the same city.
Finally, searched on RedFin.com and found two homes in the city of San Juan Capistrano.
The Trulia site allows unrestricted display of these two opportunities.
In the final analysis, the house on Cook Lane offers the best value investment.

15.e) Discussion

After exploring demographics, real estate market conditions, the competitive landscape and MLS listings, there are a few areas where further analysis would have been beneficial. For example, there is a certain level of risk associated with city planning, zoning laws and property taxes. Predictive analytics could be applied to these factors. In addition, the location in California always requires an earthquake assessment as part of the selection process. The Folium application and geospatial earthquake data could be used to determine the level of risk by geographic location. Finally, working directly with a real estate agent to get full access to MLS real estate data would provide the opportunity to conduct clustering analysis on MLS data.

16. Links to Sources and Analysis Tools

16.a) Wikipedia and Geospatial Data Sources

California GIS Data - GIS and Digital Spatial Data - Research Guides at Humboldt State University

California GIS Map & Information Sites

GIS and Mapping - Ventura County

Geospatial Data Download

California Geospatial Data - Geographic Information Systems (GIS) - LibGuides at California State University, Northridge

County View Ventura County, California
GIS and Mapping - Ventura County
Orange County, California - Geospatial Data Download
GIS Lounge - Maps and GIS
Geoportal
US Cities Database | Simplemaps.com
California, USA Map Lat Long Coordinates
Download: Zip Code Latitude Longitude City State County CSV - Gaslamp Media
Villa Park, California - Wikipedia
Rancho Santa Margarita, California - Wikipedia
Seal Beach, California - Wikipedia
Laguna Niguel, California - Wikipedia
Yorba Linda, California - Wikipedia
Camarillo, California - Wikipedia
Laguna Hills, California - Wikipedia
La Palma, California - Wikipedia
San Clemente, California - Wikipedia
Moorpark, California - Wikipedia
Tustin, California - Wikipedia
San Juan Capistrano, California - Wikipedia
Ojai, California - Wikipedia
Fountain Valley, California - Wikipedia
Placentia, California - Wikipedia
Los Alamitos, California - Wikipedia
Laguna Beach, California - Wikipedia
Mission Viejo, California - Wikipedia
IP2Location™ LITE IP-COUNTRY-REGION-CITY-LATITUDE-LONGITUDE-ZIPCODE Database | IP2Location LITE

16.b) Real Estate and Demographic Data Sources¶

Real Estate Data

Vital Signs: Home Prices – by zip code | MTC - Open Data (beta)
Vital Signs: Home Prices – by zip code | MTC - Open Data (beta)
Data Center - Redfin Real-Time
27902 Via Madrina, San Juan Capistrano, CA 92675 - 5 Bed, 6 Bath Single-Family Home - MLS #OC18268633 - 64 Photos | Trulia
32061 Cook Ln, San Juan Capistrano, CA 92675 - 6 Bed, 8 Bath Single-Family Home - MLS #PW18235467 - 48 Photos | Trulia

Demographic Data

Crime Trends in California - Public Policy Institute of California

16.c) Analysis Tools

Search - real estate | data.world
San Juan Capistrano, CA to LAX - Google Maps
Laguna Beach, California to LAX - Google Maps
Documentation - Foursquare Developer