KELVIN CORNISH

5/10/2023

DATA SOURCE

California Regional Multiple Listing Service (CRMLS) Data: Data Sourcing

I consider this to be an internal data source. I have access to CRMLS data as a Broker Participant who meets several requirements including being a licensed real estate broker in the state of California and being a member of the Pacific West Association of Realtors. The data is highly regulated, closely monitored and should be considered very accurate and reliable.

CRMLS Data: Data Collection

This is administrative data. It is submitted for input into the MLS database by licensed real estate professionals, usually the listing broker, who represents that the information is accurate to the best of their knowledge. MLS rules require initial entry of property information, closing or other disposition of the property, and other property events to be updated in the database in a timely fashion. Any information that is deemed to be inaccurate must be corrected by the participant within two days of being notified. The consequence of not making updates or corrections when instructed include monetary fines and/or loss of access to the MLS database. The data is the foundation of and the entire purpose of the multiple listing service.

CRMLS Data: Data Contents

The data contains information regarding property characteristics (location, type, size, and other amenities) and information regarding the real estate transaction (sale price, sale date, sale type, days on market, and cooperating broker commission).

CRMLS Data: Data Relevance

The objective is to determine if certain characteristics of the property and / or of the real estate transaction can be used to predict future sale values and days on market. The data contains most of the necessary information to investigate this possibility.

Median Household Income: Data Sourcing

This is external data. It is provided by the United States Census Bureau via the QuickFacts website which provides various data and information for states, counties, cities, and some towns or by zip code. It is government data and is a trustworthy data source.

Median Household Income Data: Data Collection

This is survey data. The data is collected in the American Community Survey (ACS) and the Puerto Rico Community Survey (PRCS) conducted annually by the U.S. Census Bureau. A sample of over 3.5 million housing unit addresses is interviewed each year over a 12-month period.

Median Household Income Data: Data Contents

The data is the most recent available and represents a 5-year estimate (2017 – 2021) of the median household income of the selected cities. It is presented in 2021 dollars.

Median Household Income Data: Data Relevance

The data is relevant because it contains one of the primary characteristics of interest, the median household income of the targeted cities.

DATA PROFILE

Initial Data Cleaning and Data Wrangling

Action	Columns Affected	Comment/Reason
Append files containing MLS data		**Append total of 22 individual files as exported from CRMLS Matrix website (The site has an export limit of 2000 records per export)**
Remove columns	**"Source.Name", " ", "S", and "Area"**	**Removed four columns that will not be needed in this analysis**
Remove rows containing Null		**Removed total of 223 rows that contained Null values across all columns**
Remove records with sale date prior to 1/1/2022		**I have decided to use data starting from 1/1/2022 to the present (5/3/2023). 664 records removed.**
Removed duplicates		**There was some overlap in my search queries and also some duplicate listings in the MLS data. 2335 duplicates were removed**
Change several column names	**"Sub Type" to "Property Type"**	**Column names changed for the sake of clarity.**
	"St#" to "Street #"	**Column names changed for the sake of clarity.**
	"St Name" to "Street Name"	**Column names changed for the sake of clarity.**
	"SLC" to "Sale Type"	**Column names changed for the sake of clarity.**
	"L/C Price" to "Sale Price"	**Column names changed for the sake of clarity. Only the Closing or "Sale Price" will be used in this analysis.**
	"Price Per Square Foot" to "Price Per SqFt"	**Column names changed for the sake of clarity.**
	"Sqft" to "Property Sqft"	**Column names changed for the sake of clarity.**
	"Yr Built" to "Year Built"	**Column names changed for the sake of clarity.**

	"LSqft/Ac" to "Lot SqFt"	Column names changed for the sake of clarity. Lot size in acres data will be removed from the column in a later step.
	"DOM/CDOM" to "DOM"	Column names changed for the sake of clarity. Combined DOM data will be removed from the column in a later step.
	"V" to "View?"	Column names changed for the sake of clarity.
	"PP" to "Pool?"	Column names changed for the sake of clarity.
	"BAC" to "Commission"	Column names changed for the sake of clarity.
	"Contract Status Change Date" to "Sale Date"	Column names changed for the sake of clarity. I am only concerned with the sale date; other status changes will not be included in the data.
Remove CDOM info from DOM column	DOM	The Combined DOM which includes days on market from previous listings is not relevant to this analysis.
Set negative DOM values to zero	DOM	Some properties sell before becoming active "For Sale" listings on the MLS. This results in a negative DOM. (<0.05% of records)
Remove acreage information from Lot SqFt column	Lot SqFt	The lot size in acres is not needed for this analysis
Remove info source information from the Year Built column	Year Built	The information source for the year a property was built is not needed for this analysis
Remove info source information from the Property SqFt column	Property Sqft	The information source for the properties square footage is not needed for this analysis
Separate Br (Bedrooms) into new column	Bd/Ba ---> Bedrooms	Move the number of bedrooms into its own column "Bedrooms" column for analysis purposes
Separate full bathrooms and 3/4 bathrooms into new column	Bd/Ba ---> Full Bath	Separate number of full bathrooms and insert into new column
Separate 3/4 bathrooms into new column	Bd/Ba ---> 3/4 Bath	Separate 3/4 bathrooms and insert into new column
Separate half-bathrooms into new column	Bd/Ba ---> Half Bath	Separate number of half bathrooms and insert into new column
Change columns from text to int	Bedrooms, Full Bath, Half Bath	
Add Full Bath & 3/4 Bath Columns	Full Bath & 3/4 Bath	A 3/4 Bathroom contains toilet, sink, and shower instead of bathtub. It is typically considered to be a full bathroom. Adding these columns together will create a true full bathroom total
Remove columns	Br/Ba, Full Bath(old), 3/4 Bath	Br/Ba was the original column containing all bedroom and bathroom info. The other two columns were temporary columns
Change columns from text to int	Property SqFt, Lot SqFt	
Change to full city names	City	Most city names were abbreviated, change all to full city name

Remove "*" from commission percentages	**Commission**	**Some commissions had and * attached to note special information about a particular record. That info is not relevant here so the special notation is not needed**
Fix commissions presented in dollars	**Commission**	**Some commissions were presented as a flat amount in dollars. Use this amount with the sale price to compute a commission percentage and input that in place of the dollar amount**
Create new commission level column	**Commission Level**	**Create a new column to define three commission levels, Low, Medium, and High for the purpose of analysis**
Merge income column into table	**Median Household Income**	**To provide median household income for the city where a property is located**

Data Exploration

Descriptive Statistical Analysis					
Variables	**Min**	**Max**	**Mean**	**Median**	**Standard Deviation**
Listing ID					
Property Type					
Street #					
Street Name					
City					
Sale Type					
Sale Price	$ 10,000.00	$ 10,875,000.00	$ 629,864.02	$586,000.00	$ 258,233.44
Price Per SqFt	$ 5.60	$ 8,923.50	$ 351.95	$ 341.96	$ 120.97
Bedrooms	0	13	3.4	3	1.0
Full Bath	0	13	2.2	2	0.7
Half Bath	0	5	0.3	300	0.0
Property Sqft	100.0	12,759.0	1,902.3	1,716.0	837.2
Year Built	0.0	222	37.8	35	24.7
Lot SqFt	300.0	2,567,426	14,539.7	7,500.0	52,524.9
DOM	0	546	28.89	13.00	38.9
View?					
Pool?					
Sale Date					
MLS					
Commission	0.00%	6%	2.17%	2.00%	0.36%
Commission Level					
Median Household Income	$ 46,834.00	$ 141,827.00	$ 81,145.10	$ 76,755.00	$ 17,321.62

Descriptive Statistical Analysis (Original)					
Variables	**Min**	**Max**	**Mean**	**Median**	**Standard Deviation**
Listing ID					
Property Type					
Street #					
Street Name					
City					
Sale Type					
Sale Price	$ 1,700.00	$ 11,000,000.00	$ 651,483.70	$595,000.00	$ 326,621.47
Price Per SqFt	$ 1.02	$ 8,923.50	$ 353.60	$ 340.51	$ 136.54
Bedrooms	0	17	3.48	3	0.996
Full Bath	0	13	2.25	2	0.779
Half Bath	0	5	0.31	0	0.000
Property Sqft	100	16274	1953.91	1749	898.612
Age	0	222	37.64	35	25.217
Lot SqFt	-	**333,495,360.00**	**56,440.80**	7,405.00	3,013,772.22
*The **Lot SqFt** column contains a large number of outliers at the upper range of, and many far beyond actual lot size values. I have access to the data entry form I believe the cause of most errors is due to agents entering the lot size in sq.ft. but selecting acres as the unit of measure. The system then calculates lot sqft, thinking acres was entered. This was veriy prevelent for condo and townhome property types. A second problem with these property types, they typically should not include a lot size at all. To address the proplem, I removed lot size info from all condo and townhome property types. I then spot checked and corrected lot size errors in cities and areas where I know extremely*					
DOM	0	976	37.09	17	49.61
View?					
Pool?					
Sale Date					
MLS					
Commission	0	**250.00%**	2.18%	2.00%	1.61%
*The **Commissions** column contained 3 outliers, including one at 250%. Two appeared to be input errors and were changed to 2.5% and 2.0%. The cause of the error on the third was unclear, but the value was set at 2% due to that being the median for commission values.*					
Commission Level					
Median Household Income	$ 46,834.00	$ 141,827.00	$ 81,126.89	$ 76,755.00	$ 17,566.30

LIMITATIONS AND ETHICAL CONCERNS

Data limitations

1. I still have some concerns about the accuracy of the Lot SqFt data. I believe many if not most of the errors have been corrected, but some may be hidden. I may need to find a method to systematically go through the data to find and correct any remaining outliers.

2. The data is required to have a time-dependent variable. My data set goes back approximately 16 months and can be broken down to weekly or possibly daily. I assume this will be sufficient, but I have not yet looked at the requirements of future assignments.

3. I'm not sure if the features of the data are optimized for the analysis I'll be performing. I may need a total room count instead of three separate columns for bedrooms and bathrooms. I may need flags for Property SqFt and Lot SqFt. Other adjustments may need to be made.

Ethical Concerns

I do not have any specific ethical concerns about the data. Once I finalize the data, I do plan to remove the address column which contains the house number and street as it is not necessary for the analysis.

QUESTIONS TO EXPLORE

1. Can home prices be predicted using property location and amenities?

2. Which property characteristics are most and least relevant when predicting home prices.

3. Can these characteristics be used to predict days on market?

4. Specifically, does the commission paid to the selling broker affect selling price or days on market?

5. How relevant are the property characteristics across different locations?