Vancouver House Price Analysis Report

Jan 31st, 2020

Yubo Wang

GOALS OF THE PROJECT

1. Collect house/building Data in Vancouver
2. Analyse Vancouver house market

STEPS

1. Web scraping

In order to use the website to search, we can either use the specific civic address of the building or use its parcel id. I went to Vancouver Open Data Portal to download a csv including all buildings' addresses and parcel id. After that, I used parcel id to get the building page in selenium, save the html page source and use Beautiful Soup to parse the html to get useful information including current value, previous year value, built year, area, number of bedrooms, number of bathrooms and garage.

Since the website has an anti-scraping algorithm, I need to set a random wait timer from 5 to 10 seconds after each search. To speed up the process, I launched 4 AWS EC2 Instances to run the scraping code, then store the collected data to a S3 bucket.

After scrapping, I got a dataframe in the following format

Geo Local A	lon	lat	full_add	total_value	prev_value	built_year	bedroom	bathroom	garage	area
Kensington	-123.077	49.24085	4868 HENF	$1,461,000	$1,644,000	1999	6	4	G	33 x 110 Ft
Kensington	-123.077	49.24112	4840 HENF	$1,538,000	$1,727,000	2012	4	3	G	33 x 110 Ft
Kensington	-123.077	49.24177	4780 HENF	$1,167,400	$1,372,000	1948	5	2	G	33 x 109.88 Ft
Kensington	-123.078	49.24193	4767 HENF	$2,012,000	$2,293,000	2011	7	4	G	44 x 111.39 Ft
Kensington	-123.077	49.24235	1386 31ST	$2,079,000	$2,276,000	2017	7	6	G	44 x 104.79 Ft
Kensington	-123.077	49.24131	4823 KNIG	$926,900	$1,076,400	1939	5	1		3541.73 Sq Ft

2. Data cleaning

Due to network error, for some search, there is no result returned, so I have to drop those empty rows. For current value and previous value, parse the string and convert them to integer for future use. For bedroom and bathroom, I fill null values with 0 since building with null bedroom and bathroom is commercial building or condo/apartment. For garage, in raw data 'G' means the house has a garage, so I just change 'G' to 1 and fill null with 0. Finally, for area, I split the string on space, if there is x in the string, multiply the dimension to get area, otherwise, just parse the area then convert to float.

After cleaning, subtract previous value from current value to get price change, divide price change by previous value to get price change rate, divide the current value by area to get the unit price.

3. Implment Machine Learning Model

- Using Linear Regression to predict residential house prices based on Region, built year, area, number of bedrooms, bathrooms, and garages.

  Since the variance of values is big, I decide to train separate models from each individual region to improve the accuracy. Normalize the data since the value range difference of columns is big. After training models, we have following R^2 score

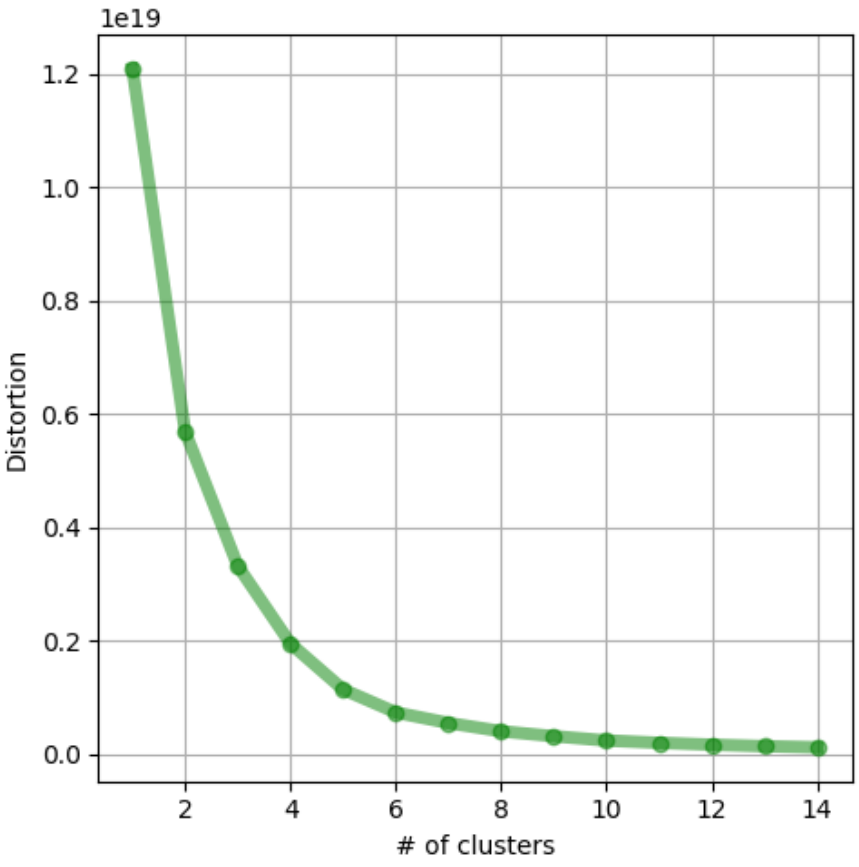
'Kensington-Cedar Cottage'	r2: 0.7275584968812627
'Renfrew-Collingwood'	r2: 0.6931773391644624
'Sunset',	r2: 0.6640833399613051
'Hastings-Sunrise'	r2: 0.7039210785582909
'Dunbar-Southlands'	r2: 0.5248994260834603
'Victoria-Fraserview'	r2: 0.7951816813279151
'Riley Park'	r2: 0.5231485037848531
'Marpole'	r2: 0.5034834962191344
'Killarney'	r2: 0.7467653870876554
'Kerrisdale'	r2: 0.6343728338318276
'Kitsilano'	r2: 0.53960635401618357
'Grandview-Woodland'	r2: 0.6680653403151944
'Arbutus-Ridge'	r2: 0.677139264661547
'Mount Pleasant'	r2: 0.47024284999609056
'Shaughnessy'	r2: 0.6471343339602749
'Oakridge'	r2: 0.46939780063189407
'West Point Grey'	r2: 0.9272779404320457
'Fairview'	r2: 0.5039335747370566
'Strathcona'	r2: 0.4826511648111762
'South Cambie'	r2: -0.449888240728034
'Downtown'	r2: -0.5364194132174658
'West End'	r2: -0.045734131990541 77

R^2 score for the last three regions is negative since the dataset after filtering of these three regions is super small. For South Cambie and West End, we only have below 50 rows. For Downtown, we only have 4 rows. So the model will not be accurate for the last three regions since the model is built for predicting house price, but there aren't many houses there.

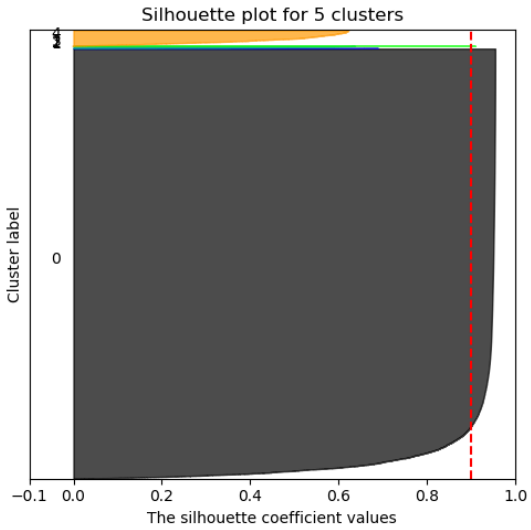
- Try to find is there specific group of buildings has high/low total value, unit price, and price change rate using Kmeans

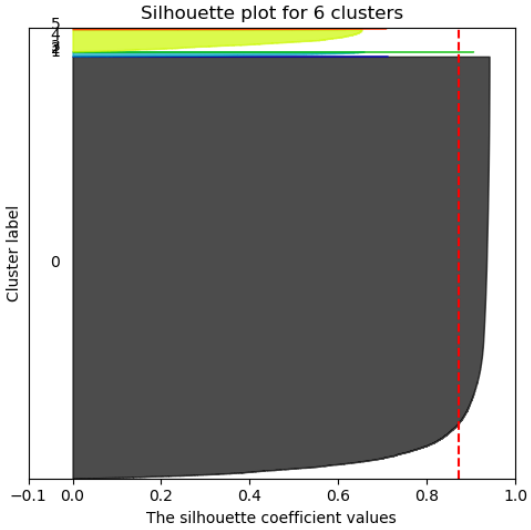
  For cluster model choices, I intended to use DBSCAN since it can handle coordinates value better, but the clustering result doesn't give any useful insights. Then I just encode the region columns then use them as location infos instead of coordinates to do the cluster in Kmeans.

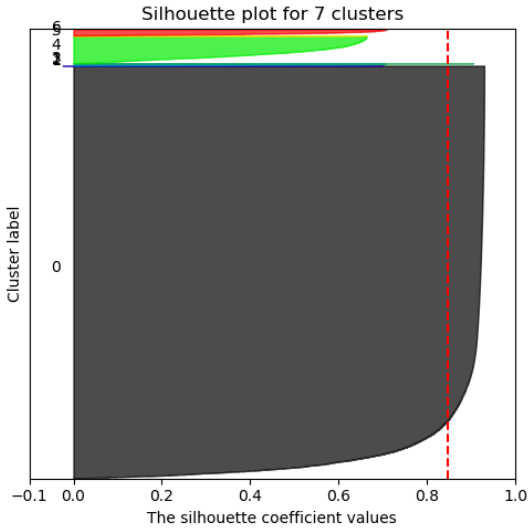
  To begin with, I use elbow graph to determine the appropriate number of clusters

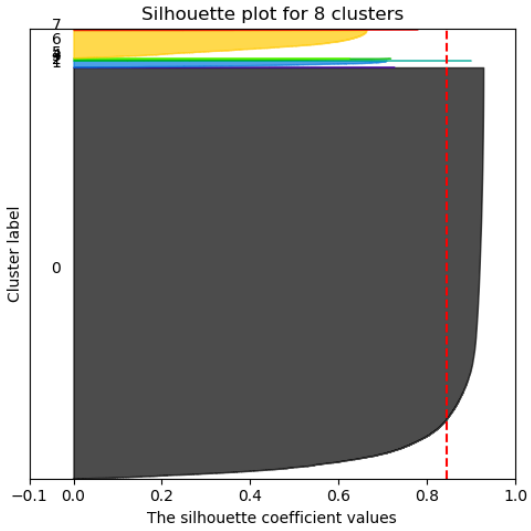


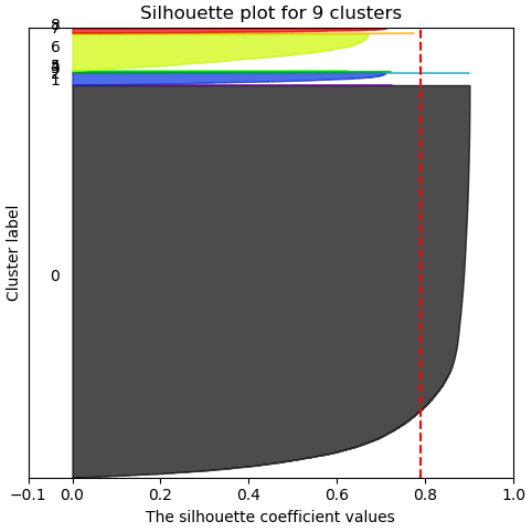
Based on the graph, I try to fit the data with 5-10 clusters and plot the silhouette graph











After exploring the cluster data, I decided to use the model with 9 clusters to get the average of each columns group by cluster.

cluster	area	bedroom	bathroom	built_year	garage	total_value	unit_price	change_rate
0	5005.13687	4.33486	3.12545	1965.03875	0.72968	1932922.91031	401.60581	-0.11198
1	161238.72293	0.00000	0.02778	1974.86111	0.00000	227369813.88889	3861.07456	0.02133
2	122553.77549	0.04255	0.02979	1968.28085	0.00426	74167467.23404	2435.27179	-0.00671
3	218846.58333	0.08333	0.04167	1980.25000	0.00000	356952375.00000	6341.49202	-0.01952
4	497333.33333	0.00000	0.00000	1981.00000	0.00000	966210600.00000	5214.82547	-0.03120
5	28124.53902	0.22398	0.30013	1964.30266	0.03807	19639203.36294	1176.71474	-0.04196
6	90336.40909	0.00000	0.00000	1971.31183	0.00000	135100792.47312	3073.01813	-0.00861
7	13412.87461	1.43554	1.72694	1964.04489	0.26730	7933987.41886	801.03808	-0.07886
8	50302.46125	0.02658	0.03987	1962.95183	0.00498	40143334.88372	1667.01187	-0.03580

Arbutus-Ridge	Downtown	Dunbar-...	Fairview	Grandview...	Hasting...	Kensington...	Kerrisdale	Killarney	Kitsilano	Marpole	Mount Pleasant	Oakridge	Renfrew...	Riley Park	Shaughnessy	South Cambie	Strathcona	Sunset	Victoria-...	West End	West Point Grey
0.03129	0.00237	0.07717	0.00570	0.03848	0.08950	0.11961	0.03858	0.05053	0.03564	0.04688	0.01928	0.02674	0.11838	0.05470	0.01869	0.01924	0.01313	0.0943...	0.07533	0.00174	0.02263
0.00000	0.55556	0.00000	0.11111	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.02778	0.00000	0.00000	0.00000	0.05556	0.00000	0.02778	0.00000	0.0000...	0.00000	0.22222	0.00000
0.02128	0.36170	0.02553	0.07660	0.01277	0.00851	0.00426	0.00426	0.02128	0.01277	0.03404	0.04681	0.00426	0.01277	0.00851	0.01277	0.02128	0.10638	0.0042...	0.00426	0.19149	0.00426
0.00000	0.83333	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.08333	0.00000	0.00000	0.00000	0.00000	0.00000	0.0000...	0.00000	0.04167	0.04167
0.00000	0.66667	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.33333	0.00000	0.00000	0.00000	0.00000	0.00000	0.0000...	0.00000	0.00000	0.00000
0.03046	0.11421	0.00571	0.13515	0.06091	0.01586	0.01904	0.05140	0.00825	0.12944	0.03744	0.11675	0.00698	0.01650	0.01396	0.02538	0.01206	0.05774	0.0114...	0.01206	0.10723	0.01206
0.00000	0.43011	0.00000	0.06452	0.00000	0.00000	0.01075	0.01075	0.01075	0.01075	0.01075	0.01075	0.00000	0.02151	0.01075	0.01075	0.01075	0.04301	0.0000...	0.00000	0.33333	0.01075
0.03027	0.05622	0.02183	0.11182	0.07455	0.01297	0.02451	0.09390	0.00680	0.08649	0.05416	0.11388	0.01627	0.01339	0.02183	0.10482	0.01483	0.05251	0.0148...	0.00577	0.03398	0.03439
0.01329	0.18106	0.00166	0.13621	0.01661	0.01993	0.01495	0.02326	0.01163	0.01661	0.04153	0.05814	0.01495	0.01661	0.01495	0.00332	0.01329	0.10465	0.0166...	0.01495	0.25581	0.00997

4. Build interactive dashboard

Build a dashboard using plotly-dash to present the analysis. Plot price, unit price and price change rate group by regions, the number of rooms, binned building areas and binned built years. Use the trained Linear Regression model in the dashboard to predict the price.

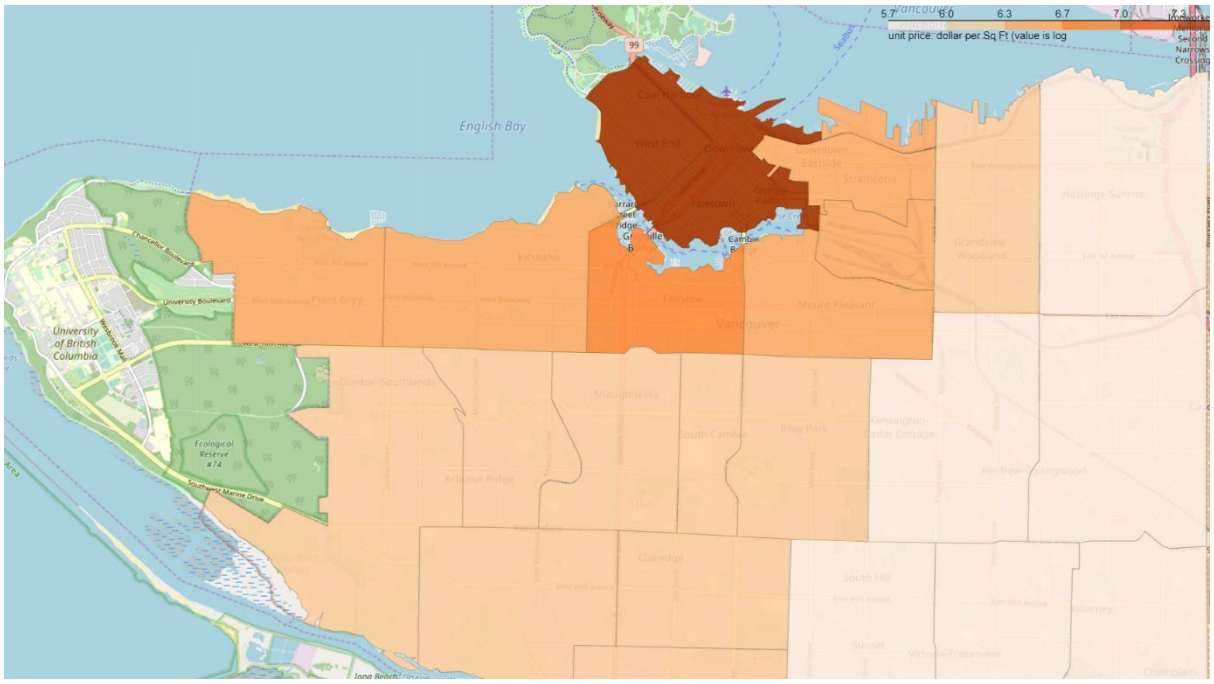
RESULT

Dashboard url: http://bit.ly/vancouver-house-price

From graphs in the dashboard, we can have several assumptions:

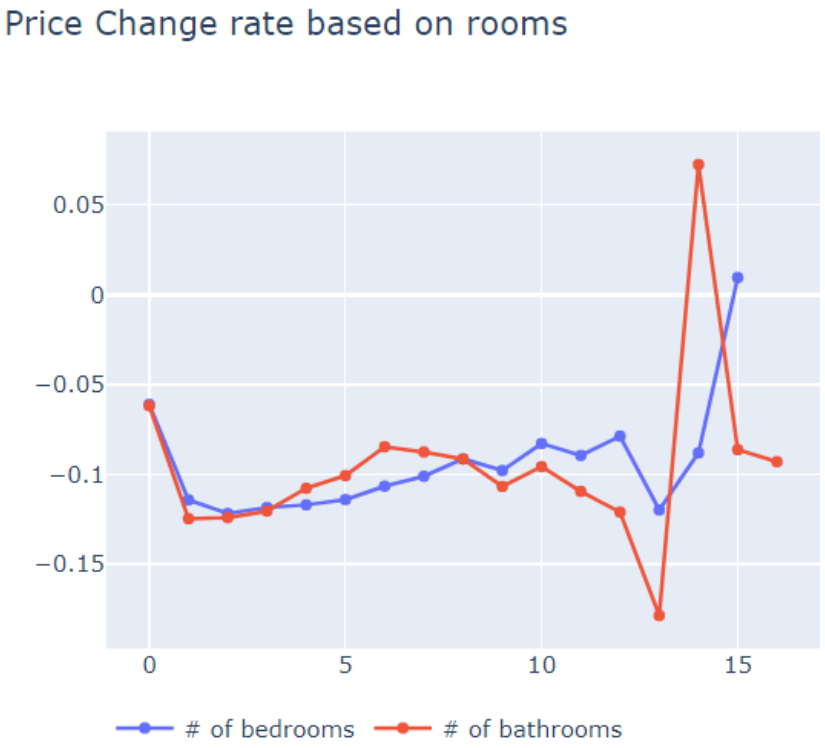
1. For average prices in a region, only prices of buildings in Downtown increases. Houses in Shaughnessy has the largest depreciation rate compred with other regions.

2. Unit price of houses in west regions is higher than unit price in east regions. Downtown has the highest unit price, and regions next to Downtown have relative higher unit prices compared with other regions. Houses in West Point Grey (top left region) has relative higher unit price because it is close to UBC

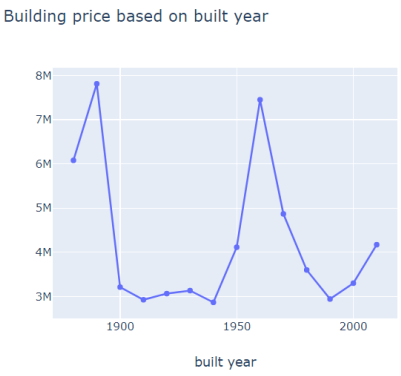


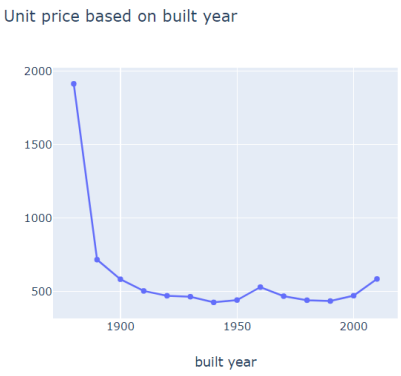
3. Unit price of houses with 6 to 8 bathrooms is higher

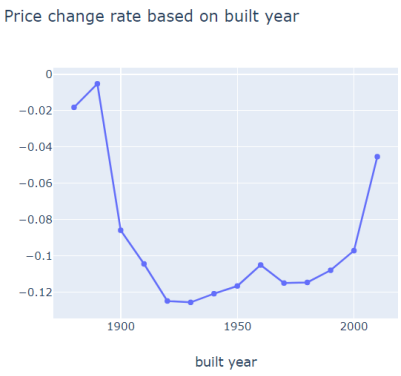
4. Houses with 6 bathrooms have the lowest depreciation rate compared with houses with other number of bathrooms.



5. Generally, price, unit price, and price change rate of buildings built before 1990 or after 2000 are higher except for building built between 1960 and 1970







From the table after clustering, we can have several assumptions:

1. Average price of houses in Vancouver is decreasing
2. Price of commercial building and condo in downtown is increasing
3. Houses in Fairview, Mount Pleasant, Shaughnessy and Kitsilano has lower depreciation rate compare with houses in other regions