Clusterización de Casas

David Rincon Morales
Noviembre 18, 2022

Problemática

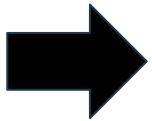
Una empresa inmobiliaria busca mejorar la segmentación de casas a través de los diferentes marketplaces disponibles. Para ello necesita clasificar cuáles son las casas “premium” dentro de la ciudad de Portland.

Ellos no tienen una forma establecida para definir qué casa es premium y cuál no, entonces podemos clasificarlas a nuestro criterio.

Lectura y Modificación de los Datos

Cada registro es una única casa (hay 39 repetidos)

- Casi 26K registros con 348 variables.
- No tenemos la variable “Premium”.
- Tenemos historia de dueños pasados y rentas pasadas de las casas.
    - Ocasiona muchos missing values.
    - Las variables categóricas pueden no estar bien homologadas.



priceHistory/2/showCountyLink	priceHistory/2/source	priceHistory/2/time	priceHistory/3/attributeSource/infoString2	priceHistory/3/date	priceHistory/3/event	price
NaN	NaN	NaN	NaN	NaN	NaN	
False	RMLS (OR)	1.613690e+12	NaN	NaN	NaN	
False	RMLS (OR)	1.619140e+12	Public Record	4/15/1996	Sold	
False	RMLS (OR)	1.619570e+12	Orion Property Management LLC	7/14/2018	Listing removed	
False	Public Record	9.936860e+11	NaN	NaN	NaN	

Lectura y Modificación de los Datos

executed in 19ms, finished 21:45:01 2022-09-30

```
'latitude',
'livingArea',
'longitude',
'lotSize',
'pageViewCount',
'postingContact',
'price',
'priceHistory',
'priceHistory',
'priceHistory',
'priceHistory',
```

Muchas son columnas rellenas de valores NaN

- Como hay muchas “categorías”, obtenemos varias columnas innecesarias.
- Eliminamos todas las columnas que no sean necesarias (por ahora) para el modelaje.
- Creamos 3 variables de relación:
  - Relación Baños - Recámaras.
  - Relación último precio de venta - tamaño de lote.
  - Relación tamaño de vivienda - tamaño de lote

Clustering Previo

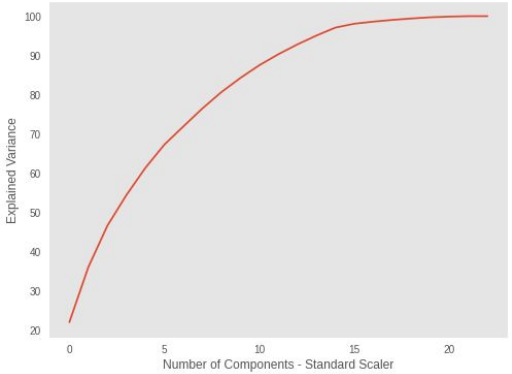
- La categorización se basó en los percentiles de diferentes variables.
- Definimos las casas como Premium si tenían al menos 5 recámaras.
  - Poco más del 10% del dataset cumplía con este objetivo.
  - Además, el precio de venta de estas casas incrementaba de manera considerable.
  - La media de precio en este conjunto era cercano a ~866K USD vs ~585K USD del total.

	25%	75%	90%	99%	mean	max
bathrooms	2.000000	3.0	4.0	5.0	2.561209e+00	23.0
bedrooms	3.000000	4.0	5.0	6.0	3.356593e+00	43.0
lastSoldPrice	390206.250000	662000.0	879996.2	1900000.0	5.850185e+05	41000000.0
p_bathrooms	0.666667	1.0	1.0	1.5	inf	inf

Se guardó en un dataframe temporal.

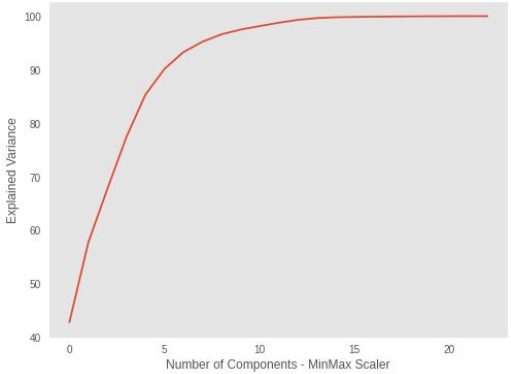
Reducción Dimensional (PCA)

Standard Scaler



Necesitamos 12 componentes para representar el 90% de los datos

MinMax Scaler

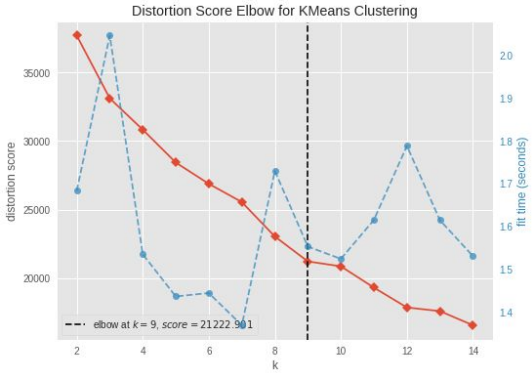


Necesitamos 6 componentes para representar el 90% de los datos

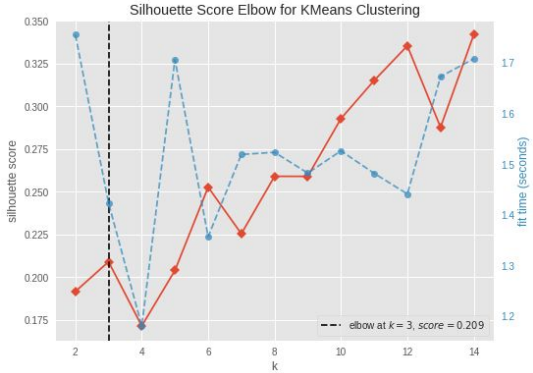
Número de Clusters

Clusters usados: 3 y 4

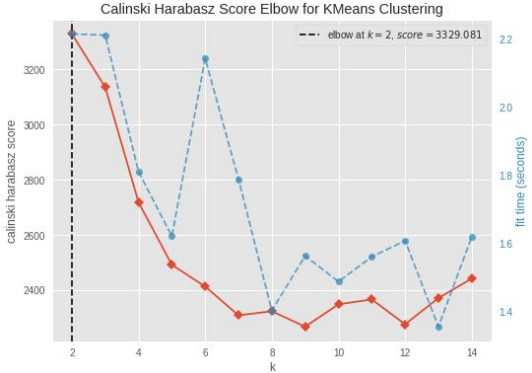
Codo



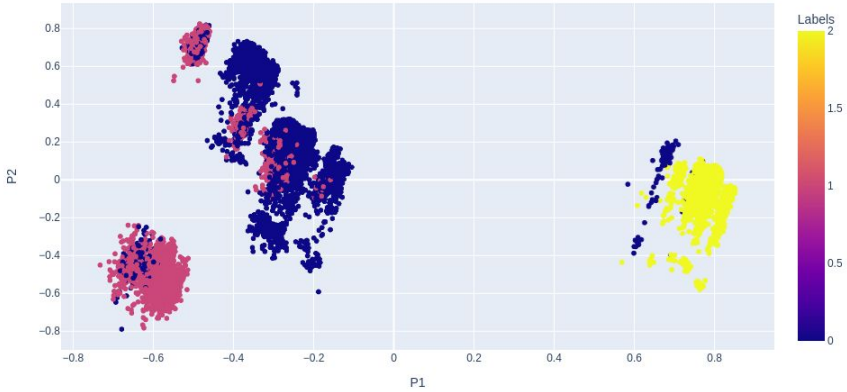
Silueta



Calinski



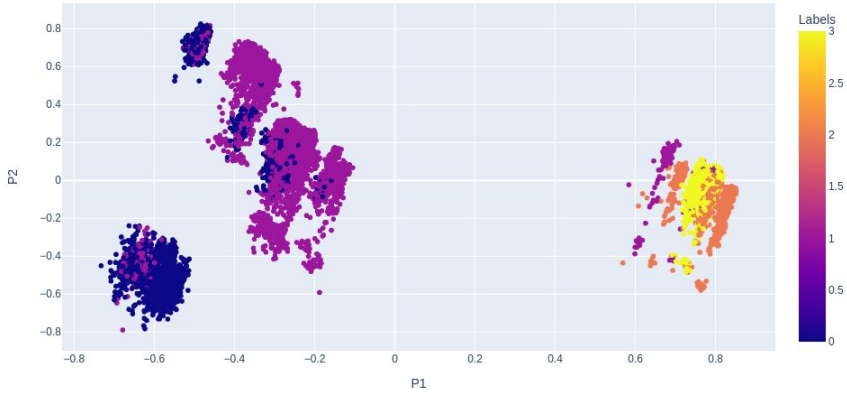
Visualización de Clusters



PCA - 3

- Es muy clara la división de clusters
- Algunos puntos del cluster 0 nos cubren a los puntos amarillos (cluster 2)

Visualización de Clusters



PCA - 4

- Aunque la división de datos es muy clara, esta cantidad de clusters nos empieza a mezclar muchos datos

Perfilamiento Ganador - PCA 3 Clusters

Nombre	Descripción
Cluster 0: Casas Promedio	• Son las casas más viejas, fueron construidas en los 1950's. • Tienen el sqft más caro con 109 USD. • Tienen el mejor potencial de adaptación a energía eléctrica
Cluster 1: Casas Premium	• Son las casas más costosas. Su precio promedio se encuentra cerca de 625K USD. • Su renta es la más cara, con un precio de 3169 USD mensuales (10% extra que las casas promedio) • Su tamaño de terreno es más del doble que las demás.
Cluster 2: Casas "Entry Level"	• Son las casas más baratas, tienen un precio promedio de 525K USD (comparado con 600k+ de las casas promedio). • Son las más nuevas, fueron construídas en los 1990's. • Su cantidad de baños y recámaras tiende a ser el mismo