Analiza corespondentelor simpla si Two-Step Cluster

Am lucrat pe baza de date " London Housing Market Analysis: Pricing Trends & F" descarcata de pe platforma Kaggle. Aceasta contine date referitoare la locuintele din Londra precum: cartier, suprafata (m^2), tipul de proprietate (apartament, casa, duplex) si pret (£).

I. Analiza corespondentelor simpla

In analiza corespondentelor simpla, se includ doua variabile categoriale care au o legatura semnificativa statistic. Asadar, am realizat testul Chi-square din meniul Crosstabs pentru variabilele cartier si pret. Variabila pret era exprimata in valoarea absoluta in lire, dar am realizat o clusterizare de tip k-means pentru a face gruparea locuintelor pe categoriile: Affordable (pret accesibil), Standard (pret mediu) si Luxury (pret ridicat).

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	178,843[a]	18	,000
Likelihood Ratio	189,633	18	,000
N of Valid Cases	1000		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 16,29.

H_0: nu exista o legatura semnificativa intre pretul locuintei si cartierul in care se situeaza aceasta

H_1: exista o legatura semnificativa intre cele doua variabile

Pentru o valoare Asymptotic Significance < 0.05 => H_0 se respinge => exista o legatura semnificativa statistic intre variabile.

Am realizat analiza corespondentelor simpla folosind variabilele cartier si pret. Am transformat valorile variabilei cartier in numere de la 1-10, iar ale variabilei pret in numere de la 1-3.

Correspondence Table

Neighborhood_nou	Price_nou			
	Affordable	Standard	Luxury	Active Margin
Camden	49	50	7	106
Chelsea	21	32	41	94
Greenwich	68	29	0	97
Islington	55	36	6	97
Kensington	30	47	37	114
Marylebone	47	47	19	113
Notting Hill	36	43	17	96
Shoreditch	60	26	3	89
Soho	41	41	14	96
Westminster	33	26	39	98
Active Margin	440	377	183	1000

Putem observa ca 68 din locuintele din cartierul Greenwich au un pret accesibil, dar nicio locuinta din acel cartier nu are un pret ridicat. Cartierul Chelsea are 41 de locuinte cu pret ridicat si doar 21 de locuinte cu un pret accesibil.

Row Profiles

Neighborhood_nou	Price_nou			
	Affordable	Standard	Luxury	Active Margin
Camden	,462	,472	,066	1,000
Chelsea	,223	,340	,436	1,000
Greenwich	,701	,299	,000	1,000
Islington	,567	,371	,062	1,000
Kensington	,263	,412	,325	1,000
Marylebone	,416	,416	,168	1,000
Notting Hill	,375	,448	,177	1,000
Shoreditch	,674	,292	,034	1,000
Soho	,427	,427	,146	1,000
Westminster	,337	,265	,398	1,000
Mass	,440	,377	,183	

Se poate observa ca 70.1% din totalul locuintelor din cartierul Greenwich incluse in esantion au un pret accesibil, iar restul de 29,9% au un pret mediu. Din totalul locuintelor din cartierul Chelsea 43.6% au un pret ridicat si doar 22.3% au un pret accesibil.

Column Profiles

Neighborhood_nou	Price_nou			
	Affordable	Standard	Luxury	Mass
Camden	,111	,133	,038	,106
Chelsea	,048	,085	,224	,094
Greenwich	,155	,077	,000	,097
Islington	,125	,095	,033	,097
Kensington	,068	,125	,202	,114
Marylebone	,107	,125	,104	,113
Notting Hill	,082	,114	,093	,096
Shoreditch	,136	,069	,016	,089
Soho	,093	,109	,077	,096
Westminster	,075	,069	,213	,098
Active Margin	1,000	1,000	1,000	

Conform tabelului, 15,5% din locuintele ce un pret accesibil sunt in cartierul Greenwich si 13,6% sunt in cartierul Shoreditch.

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	,399	,159			,891	,891	,027	,198
2	,139	,019			,109	1,000	,031	
Total		,179	178,843	,000[a]	1,000	1,000		

a. 18 degrees of freedom

Tabelul Summary indica numarul de dimensiuni de care este nevoie pentru a putea determina corespondenta dintre starile celor doua variabile, sau, cu alte cuvinte, care e cel mai probabil pret al unei locuinte in functie de cartierul in care se situeaza. Astfel, observ ca primei dimensiuni ii este atribuita o proportie din inertie de 0.891, celei de-a doua o proportie de 0.109. Cea de a doua dimensiune nu contribuie intr-o masura semnificativa la identificarea corespondentelor.

Overview Row Points[a]

Neighborhood_nou	Mass	Score in Dimension 1	Score in Dimension 2	Inertia	Contribution: Of Point to Inertia of Dimension 1	Contribution: Of Point to Inertia of Dimension 2	Contribution: Of Dimension to Inertia of Point 1	Contribution: Of Dimension to Inertia of Point 2	Contribution: Of Dimension to Inertia of Point Total
Camden	,106	-,395	,518	,011	,041	,204	,624	,376	1,000
Chelsea	,094	1,068	-,188	,043	,269	,024	,989	,011	1,000
Greenwich	,097	-,905	-,443	,034	,199	,136	,923	,077	1,000
Islington	,097	-,541	-,039	,011	,071	,001	,998	,002	1,000
Kensington	,114	,668	,204	,021	,127	,034	,969	,031	1,000
Marylebone	,113	-,016	,215	,001	,000	,037	,016	,984	1,000
Notting Hill	,096	,064	,393	,002	,001	,106	,070	,930	1,000
Shoreditch	,089	-,765	-,479	,024	,130	,146	,880	,120	1,000
Soho	,096	-,100	,275	,001	,002	,052	,274	,726	1,000
Westminster	,098	,805	-,607	,030	,159	,259	,834	,166	1,000
Active Total	1,000			,179	1,000	1,000			

a. Symmetrical normalization

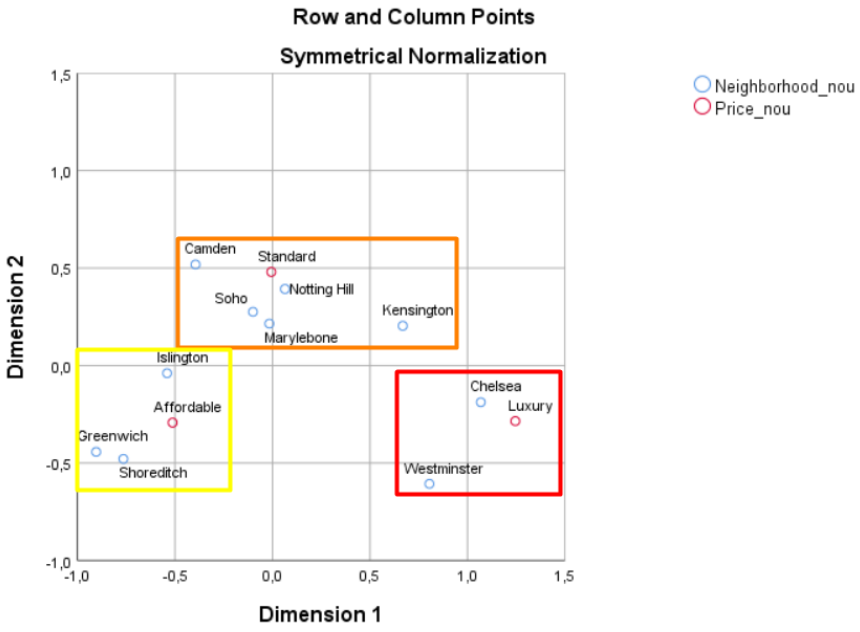
Observ ca locuintele din cartierul Chelsea contribuie la inertia din prima dimensiune in proportie de 26.9%, respectiv 2.4% in a doua dimensiune. Cartierul Westminster contribuie la inertia din prima dimensiune in proportie de 15.9%, respectiv 25.9% in a doua dimensiune. Cartierele Islington, Marylebone, Soho nu contribuie semnificativ la nicio dimensiune. Starile Greenwich si Shoreditch contribuie in proportii similare la exprimarea inertiei din prima dimensiune (19.9% respectiv 13%), deci cele doua puncte vor fi relativ apropiate una de cealalta pe grafic.

Overview Column Points[a]

Price_nou	Mass	Score in Dimension 1	Score in Dimension 2	Inertia	Contribution: Of Point to Inertia of Dimension 1	Contribution: Of Point to Inertia of Dimension 2	Contribution: Of Dimension to Inertia of Point 1	Contribution: Of Dimension to Inertia of Point 2	Contribution: Of Dimension to Inertia of Point Total
Affordable	,440	-,512	-,293	,051	,289	,271	,898	,102	1,000
Standard	,377	-,006	,480	,012	,000	,623	,000	1,000	1,000
Luxury	,183	1,245	-,285	,115	,711	,106	,982	,018	1,000
Active Total	1,000			,179	1,000	1,000			

a. Symmetrical normalization

Aceleasi contributii sunt calculate si pentru starile variabilei pret. Astfel, observ ca Luxury contribuie la exprimarea inertiei din prima dimensiune in proportie de 71.1% si in cea de-a 2- a dimensiune in proportie de 10.6%.



Toate aceste lucruri pot fi mai bine intelese pe baza graficului realizat cu contributiile fiecarei stari din ambele variabile. Cartierele in care este cel mai probabil sa se regasesca locuinte cu preturi accesibilie sunt: Islington, Greenwich si Shoreditch, deoarece punctele lor pe grafic sunt apropiate. Cartierele in care se regasesc locuinte cu preturi ridicate sunt: Chelsea si Westminster.

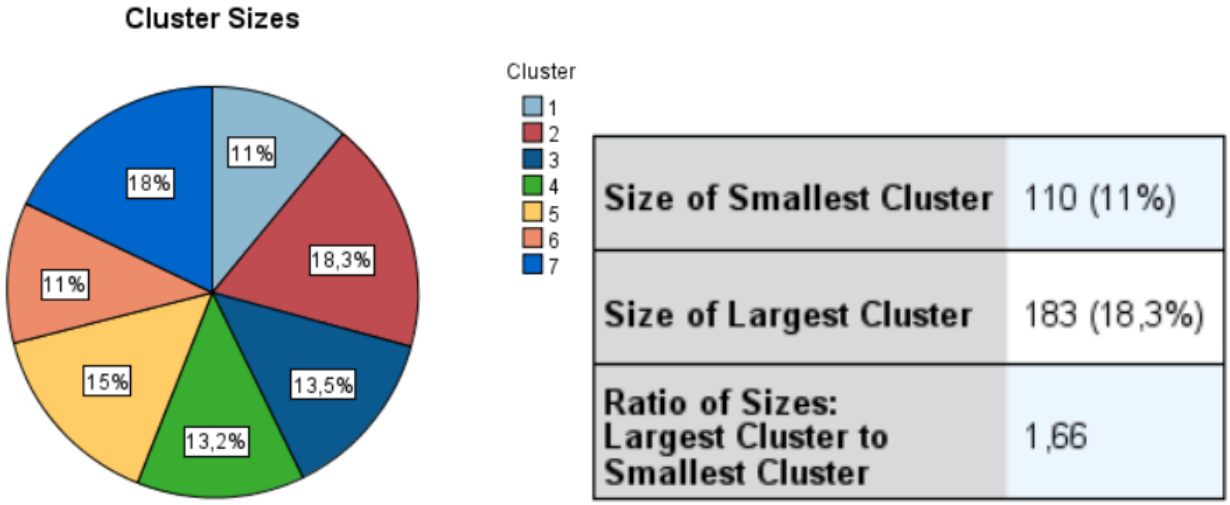
II. Two-Step Cluster

Pentru realizarea acestei analize am folosit aceeasi baza de date. In lista Continuous Variables am introdus variabila suprafata (m^2) si in lista Categorical Variables am introdus variabilele: tipul de proprietate si pretul (Affordable, Standard, Luxury).

Model Summary

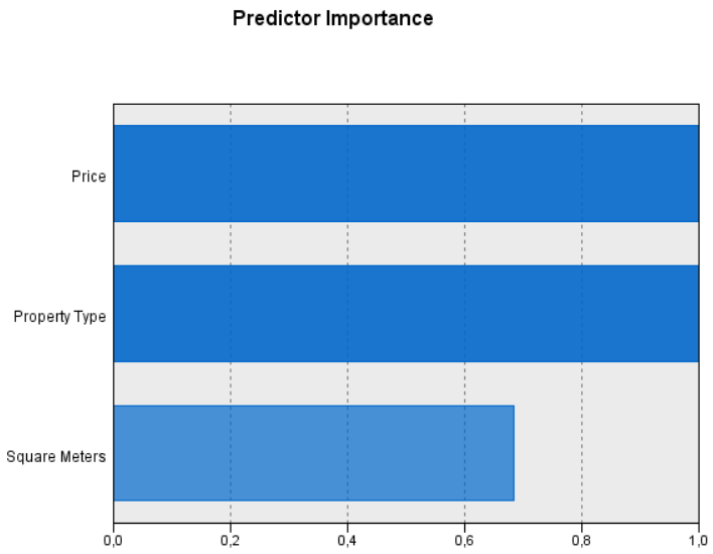
Algorithm	TwoStep
Inputs	3
Clusters	7

Desi analiza a generat automat 8 clustere, am ales sa limitez numarul la 7, aceasta fiind cea mai simpla structura care mentine calitatea clusterelor. In cazul meu, analiza depaseste valoare de 0.5 si se incadreaza in zona marcata cu verde, Good. Asadar, analiza realizata este de calitate si valida.



Size of Smallest Cluster	110 (11%)
Size of Largest Cluster	183 (18,3%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1,66

Cele 7 clustere sunt aproximativ egale ca marime, ceea ce era de dorit. In clusterul 2 au fost repartizate 18.3% dintre locuintele incluse in esantion.



Un lucru folositor de urmarit este importanta predictorilor pe baza carora s-a facut impartirea in grupe.Variabilele pret si tipul de proprietate au cele mai mari ponderi in procesul de repartizare, urmate de variabila suprafata(m^2) cu o pondere de 68%.

Impartirea propriu-zisa in clustere se poate vedea din tabelul urmator:

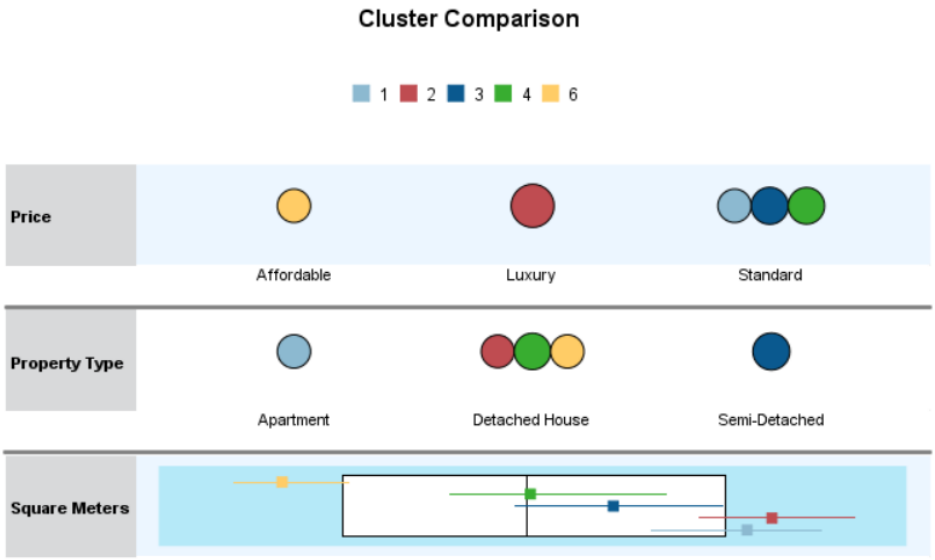
Clusters

Input (Predictor) Importance
1,0 0,8 0,6 0,4 0,2 0,0

Cluster	1	2	3	4	5	6	7
Label							
Description							
Size	11,0% (110)	18,3% (183)	13,5% (135)	13,2% (132)	15,0% (150)	11,0% (110)	18,0% (180)
Inputs	Price Standard (100,0%)	Price Luxury (100,0%)	Price Standard (100,0%)	Price Standard (100,0%)	Price Affordable (100,0%)	Price Affordable (100,0%)	Price Affordable (100,0%)
	Property Type Apartment (100,0%)	Property Type	Property Type	Property Type	Property Type	Property Type	Property Type Apartment (100,0%)
	Square Meters 200,57	Square Meters 211,81	Square Meters 173,58	Square Meters 155,62	Square Meters 101,07	Square Meters 85,83	Square Meters 112,37

In clusterul 6 au fost clasate locuintele cu cea mai mica suprafata si cu un pret accesibil. In clusterul 2 au fost clasate locuintele cu cea mai mare suprafata si cu un pret ridicat.

Aceste lucruri sunt reprezentate si grafic, prin intermediul box-plot-urilor respectiv o scala liniara:



- In clusterul 6 s-au clasat casele cu un pret accesibil si o suprafata sub medie.
- In clusterul 4 s-au clasat casele cu un pret standard si o suprafata medie.
- In clusterul 2 s-au clasat casele cu un pret ridicat si o suprafata peste medie.