UNIVERSITATEA BABEȘ-BOLYAI

FACULTATEA DE ȘTIINȚE ECONOMICE ȘI GESTIUNEA AFACERILOR

CLUJ-NAPOCA

PROIECT

ANALIZA DATELOR

Bonaț Paula-Mihaela

Statistică și Previziune Economică, anul III

Cuprins

Analiza simplă a corespondențelor

Scopul acestei analize este de a descrie legăturile sau corespondenţele dintre două variabile sau două grupe de variabile, respectiv de a studia simultan liniile şi coloanele unui tabel de contingenţă, pentru a descrie structura liniilor în funcţie de legătura lor cu coloanele sau invers, structura coloanelor în funcţie de legătura lor cu liniile tabelului analizat.

Pentru această analiză am folosit variabila *var00106* care reprezintă *Venitul familiei* și variabila *var00101* care reprezintă *4) Care este super/hypermarketul la care mergeți cel mai des?* din baza de date *Bila(1).sav*.

Calea metodei este: Analyze – Dimension Reduction – Correspondence Analysis. În căsuța apărută trecem la Row: var00106, definind intervalul răspunsurilor de la 0 la 8 și la Column: var00101, definind intervalul răspunsurilor de la 0 la 4. De la meniul Statistics mai bifăm Row Profiles și Column Profiles, iar de la Plots mai bifăm Row Points și Column Points.

Primul tabel este cel de corespondență care reflectă distribuția intervievaților în raport cu cele două variabile. Dintre familiile cu un venit sub 300 RON 2 merg cel mai des la Bila să se aprovizioneze și niciuna la Selgros. Dintre familiile cu venitul între 500-900 de RON 13 preferă tot Bila pentru aprovizionare. Dintre familiile cu venit între 900-1500 RON aleg să se aprovizioneze cel mai des din Kaufland și Bila. Dintre familiile cu venit peste 4000 RON niciuna nu preferă Kaufland. Avem și o categorie de familii care nu a dorit să dezvăluie venitul familiei, însă merg cel mai des la Bila, Selgros și Metro.

Correspondence Table

	4). Care este super/hypermarketul la care mergeti cel mai des?				
19) Venitul familiei?	Bila	Selgros	Metro	Kaufland	Active Margin
< 300 RON	2	0	1	2	5
[3 00 - 500]RON	2	1	0	5	8
(500 - 900] RON	13	10	3	8	34
(900 - 1500] RONl	24	16	12	25	77
(1500- 2500] RON	26	25	8	18	77
(2500-4000] RON	10	6	13	8	37
> 4000 RON	2	3	10	0	15
Confidential	2	2	2	1	7
Active Margin	81	63	49	67	260

Având ca punct de pornire aceste date se vor calcula frecvenţele condiţionate pentru fiecare variabilă în parte. Astfel avem aceste frecvenţe pentru prima variabilă, denumite şi ‘’profilele’’ liniilor. Așadar, familiile cu un venit sub 300 RON aleg Bila și Selgros ca și loc pentru aprovizionare în proporție de 40% și Metro în proporție de 20%. Familiile cu un venit între 1500-2500 RON merg în proporție de 33.8% la Bila și doar în proporție de 10.4% la Metro. Familiile cu un venit peste 4000 RON merg în proporție de 66.7% la Metro și 0% la Kaufland.

Row Profiles

	4). Care este super/hypermarketul la care mergeti cel mai des?				
19) Venitul familiei?	Bila	Selgros	Metro	Kaufland	Active Margin
< 300 RON	.400	.000	.200	.400	1.000
[3 00 - 500]RON	.250	.125	.000	.625	1.000
(500 - 900] RON	.382	.294	.088	.235	1.000
(900 - 1500] RONl	.312	.208	.156	.325	1.000
(1500- 2500] RON	.338	.325	.104	.234	1.000
(2500-4000] RON	.270	.162	.351	.216	1.000
> 4000 RON	.133	.200	.667	.000	1.000
Confidential	.286	.286	.286	.143	1.000
Mass	.312	.242	.188	.258	

În următorul tabel avem frecvențele pentru cea de-a doua variabilă, denumite și ”profilele” coloanelor. După cum se poate vedea, Bila este ales de familiile cu un venit între 1500-2500 RON cu o proporție de 32.1%, Selgros este ales de familiile cu un venit între 1500-2500 RON cu o proporție de 39.7%, Metro este ales de familiile cu un venit între 2500-4000 RON cu o proporție de 26.5% și Kaufland este ales de familiile cu venitul între 900-1500 RON în proporție de 37.3%.

Column Profiles

	4). Care este super/hypermarketul la care mergeti cel mai des?				
19) Venitul familiei?	Bila	Selgros	Metro	Kaufland	Mass
< 300 RON	.025	.000	.020	.030	.019
[3 00 - 500]RON	.025	.016	.000	.075	.031
(500 - 900] RON	.160	.159	.061	.119	.131
(900 - 1500] RONl	.296	.254	.245	.373	.296
(1500- 2500] RON	.321	.397	.163	.269	.296
(2500-4000] RON	.123	.095	.265	.119	.142
> 4000 RON	.025	.048	.204	.000	.058
Confidential	.025	.032	.041	.015	.027
Active Margin	1.000	1.000	1.000	1.000	

În cele ce urmează se vor calcula valorile proprii, vectorii proprii şi dimensiunile pornind de la tabelul profilelor coloanelor deoarece numărul acestora este mai mic decât al liniilor. Astfel rezultă trei dimensiuni ce vor fi reţinute în analiză. Astfel, observăm ca primei dimensiuni ii este atribuită o proporție din inerție de 0.777, celei de-a doua o proportie de 0.201 iar celei de-a treia o proportie de 0.021. Cum cea de a treia dimensiune nu contribuie într-o măsură semnificativă la identificarea corespondențelor, numarul optim de dimensiuni este 2. Cele 2 dimensiuni reprezintă 97.9% din inerție.

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	.385	.148			.777	.777	.066	.008
2	.196	.038			.201	.979	.061	
3	.064	.004			.021	1.000		
Total		.190	49.480	.000[a]	1.000	1.000		

a. 21 degrees of freedom

Următorul tabel arată cantitatea de informație din fiecare dimensiune dată de stările variabilei ”Venitul familiei”. Astfel, putem afirma că 61.2% din informațiile dimensiunii 1 sunt date de familiile al căror venit este peste 4000 RON. De altfel, 58.1% din informațiile aduse de familiile al căror venit este cuprins între 900-1500 RON se regăsesc în dimensiunea 2.

Overview Row Points[a]

19) Venitul familiei?	Mass	Score in Dimension		Inertia	Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
< 300 RON	.019	-.085	-1.163	.007	.000	.133	.008	.763	.771
[3 00 - 500]RON	.031	-.980	-1.381	.024	.077	.300	.473	.478	.951
(500 - 900] RON	.131	-.370	.358	.011	.046	.086	.638	.305	.943
(900 - 1500] RONl	.296	-.174	-.288	.008	.023	.126	.418	.581	.999
(1500- 2500] RON	.296	-.302	.420	.021	.070	.268	.498	.492	.990
(2500-4000] RON	.142	.648	-.286	.026	.156	.060	.901	.089	.991
> 4000 RON	.058	2.021	.155	.091	.612	.007	.994	.003	.997
Confidential	.027	.458	.395	.003	.015	.021	.725	.274	.999
Active Total	1.000			.190	1.000	1.000			

a. Symmetrical normalization

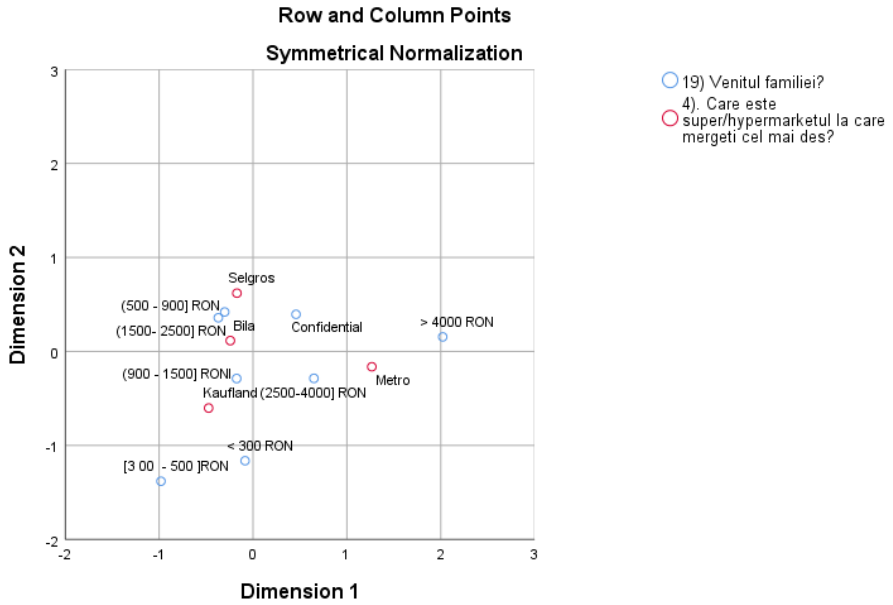
Următorul tabel arată cantitatea de informație din fiecare dimensiune dată de stările variabilei ”Care este super/hypermarketul la care mergeți cel mai des”. Așadar, 78.5% din informațiile dimensiunii 1 sunt date de starea ”Metro”. 87.7% din informația adusă de starea ”Selgros” se regăsește în dimensiunea 2.

Overview Column Points[a]

4). Care este super/hypermarketul la care mergeti cel mai des?	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 Total
Bila	.312	-.242	.115	.010	.047	.021	.678	.078	.756
Selgros	.242	-.171	.620	.022	.018	.476	.124	.827	.951
Metro	.188	1.266	-.163	.117	.785	.026	.992	.008	1.000
Kaufland	.258	-.472	-.602	.041	.150	.478	.541	.447	.989
Active Total	1.000			.190	1.000	1.000			

a. Symmetrical normalization

Graficul ”Row and Column Points” redă apropierea ambelor variabile pe cele două dimensiuni.



Aici am putea spune că familiile al căror venit este cuprins între 500-900 și 1500-2500 RON frecventează Selgros și Bila pentru a se aproviziona, iar persoanele care își fac cumpărăturile de la Kaufland au un venit cuprins între 900-1500 RON.

Analiza componentelor principale

Analiza componentelor principale are ca obiectiv prezentarea sintetică a unui tabel de date în care unităţile sunt descrise prin multiple variabile cantitative. Această descriere trebuie să permită:

- sinteză a informaţiei, variabilele descriptive sunt regrupate în factori sintetici, denumiţi componente principale, astfel încât pierderea de informaţie să fie minimă;
- poziţionarea unităţilor prin raportare la componentele principale ceea ce va permite punerea în evidenţă de tipuri de unităţi.

Problema analizei componentelor principale constă în a reduce cele p variabile iniţiale într-un număr de q variabile denumite "componente principale" sau factori, q<p.

Pentru această metodă am folosit baza de date ”baza 1.sav” unde se regăsesc un set de întrebări și răspunsuri cu privire la jocurile de noroc. Eu am luat în considerare întrebarea privitoare la efectele jocurilor de noroc (numărul 24). Întrebarea este: În ce măsură jocurile de noroc ți-au afectat următoarele aspecte, iar răspunsurile sunt: viața personală, relațiile cu prietenii, situația școlară, situația financiară, situația familială, timpul liber. Scala de măsurare este de la 1 la 5, 1 însemnând dezacord total, iar 5 însemnând acord total.

Primul pas în analiză este acela de a testa calitatea datelor. Pentru testare se foloseste indicatorul Cronbach’s Alpha. Calea pentru testare este Analyze -> Scale -> Reliability Analysis

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.887	.897	6

Indicatorul Cronbachs Alpha indică o valoare de 0.887. Intervalul recomandat pentru calitatea datelor este 0.7-0.9. Indicatorul nostru obținut aparține intervalului, deci putem afirma că datele au o calitate ridicată și putem continua analiza.

Tabelul ”Item Statistics” indică statisticile descriptive (media și abaterea) ale variabilelor, totalul observațiilor fiind 134.

Item Statistics

	Mean	Std. Deviation	N
24) Viata personala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	4.12	1.151	134
24) Relatiile cu prietenii (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	4.46	.873	134
24) Situatia scolara/profesionala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	4.52	.963	134
24) Situatia financiara (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	3.59	1.270	134
24) Situatia familiala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	4.54	.898	134
24) Timpul liber (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	3.75	1.102	134

În continuare, ne uităm la matricea corelațiilor:

Inter-Item Correlation Matrix

	24) Viata personala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	24) Relatiile cu prietenii (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	24) Situatia scolara/profesionala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	24) Situatia financiara (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	24) Situatia familiala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	24) Timpul liber (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)
24) Viata personala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	1.000	.753	.615	.693	.738	.487
24) Relatiile cu prietenii (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.753	1.000	.667	.587	.851	.444

24) Situatia scolara/profesionala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.615	.667	1.000	.515	.759	.353
24) Situatia financiara (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.693	.587	.515	1.000	.557	.393
24) Situatia familiala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.738	.851	.759	.557	1.000	.458
24) Timpul liber (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.487	.444	.353	.393	.458	1.000

Matricea corelațiilor este o matrice pătratică și simetrică cu 6 linii și 6 coloane. Aici putem observa corelațiile dintre variabile. Ceea ce ne interesează este ca toate să aibă același semn, adică ori toate să fie negative, ori toate să fie pozitive. Datele noastre au toate corelații pozitive, adică între variabile există o legătură directă. Pe diagonală corelațiile au valoarea 1, deoarece e calculată corelația între aceeași variabilă, iar aportul de informație este adus în egală măsură de toate cele 6.

Ca și interpretare am putea zice că între viața personală și relațiile cu prietenii în urma efectelor participării la jocurile de noroc există o corelație de 0.753, corelație pozitivă și de intensitate mare. Acest lucru înseamnă că o creștere într-una din variabile va aduce cu sine creșterea și în cealaltă variabilă, asemănător și în cazul descreșterii.

Metoda începe acum. Calea metodei este: Dimension Reduction – Factor și ducem cele 6 variabile în căsuța Variables. Aici selectăm:

- Descriptives: KMO and Bartlett s test of sphericity
- Extract: Fixed number of factors: 1 (așa mi-am propus), Scree plot
- Scores: Save as variables, Display factor score coefficient matrix.

Tabelul KMO:

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.860
Bartlett's Test of Sphericity	Approx. Chi-Square	530.240
	df	15
	Sig.	.000

KMO este folosit pentru măsurarea calității globale a analizei. Formulăm ipoteza nulă: Statistica KMO = 0, adică nu există corelație globală între itemi. Alternativa formulată este aceea că diferă semnificativ de 0, adică ne așteptăm la o analiză de bună calitate. Valoarea estimată este de 0.86 și este mai mare decât 0.5. Sig = 0.000 și indică probabilitatea de acceptare a ipotezei false când ea e adevărată. Așadar, ipoteza nulă se respinge și se acceptă alternativa, adică analiza este adecvată.

Tabelul Communalities ne arată ce proporție din variație a fost reținută din fiecare variabilă

Communalities

	Initial	Extraction
24) Viata personala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	1.000	.783
24) Relatiile cu prietenii (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	1.000	.801
24) Situatia scolara/profesionala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	1.000	.655
24) Situatia financiara (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	1.000	.582
24) Situatia familiala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	1.000	.823
24) Timpul liber (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	1.000	.368

Extraction Method: Principal Component Analysis.

Astfel, se poate observa că variabila ”situyația familială” a fost cel mai bine surprinsă în analiză, în proporție de 82.3%, în timp ce variabila ”timpul liber” a fost surprinsă într-o proporție de doar 36.8%.

În tabelul ”Total Variance Explained” sunt afișate valorile propri calculate

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.011	66.856	66.856	4.011	66.856	66.856
2	.717	11.956	78.812			
3	.557	9.291	88.103			
4	.355	5.921	94.024			
5	.227	3.784	97.808			
6	.132	2.192	100.000			

Extraction Method: Principal Component Analysis.

Acestea au fost ordonate descrescător de la 4 la 0.1. Suma urmei matricii este 6. Dacă insumăm aceste valori propri vom obține tot 6, însă proporțiile sunt diferite în acest caz din cauza vectorilor propri. Informația este adusă diferit, așadar prima compoonentă preia informația 4 din 6, adică 66.85% din 100%, a doua componentă preia 0.7 din 6,a dică 11.95% din 100% și așa mai departe. Prima componentă reprezintă componenta latentă și aduce 66.85% din informația inițială.

Tabelul ”Component Matrix” arată coeficienții de corelație dintre itemii inițiali și variabila latentă.

Component Matrix[a]

	Component 1
24) Viata personala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.885
24) Relatiile cu prietenii (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.895
24) Situatia scolara/profesionala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.809
24) Situatia financiara (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.763
24) Situatia familiala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.907
24) Timpul liber (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.607

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

După cum putem observa, corelația dintre variabila latentă și variabila ”situație familială” este în proporție de 90%, indicând o intensitate mare. Corelația dintre variabila latentă și variabila ”timpul liber” este în proporție de 60.7%, indicând o intensitate medie. Corelațiile diferită, datorită faptului că inensitățile relațiilor sunt diferite.

Tabelul ”Component Score Coefficient Matrix” arată coeficienții prin care se leagă itemii inițiali de componenta latenă

Component Score Coefficient Matrix

	Component 1
24) Viata personala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.221
24) Relatiile cu prietenii (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.223
24) Situatia scolara/profesionala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.202
24) Situatia financiara (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.190
24) Situatia familiala (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.226
24) Timpul liber (In ce masura jocurile de noroc ti-au afectat urmatoarele aspecte:)	.151

Extraction Method: Principal Component Analysis.
Component Scores.

*Componenta 1 = 0.221 * variabila 1 + 0.223 * variabila 2 + 0.202 * variabila 3 + 0.190 * variabila 4 + 0.226 * variabila 5 + 0.151 * variabila 6.*

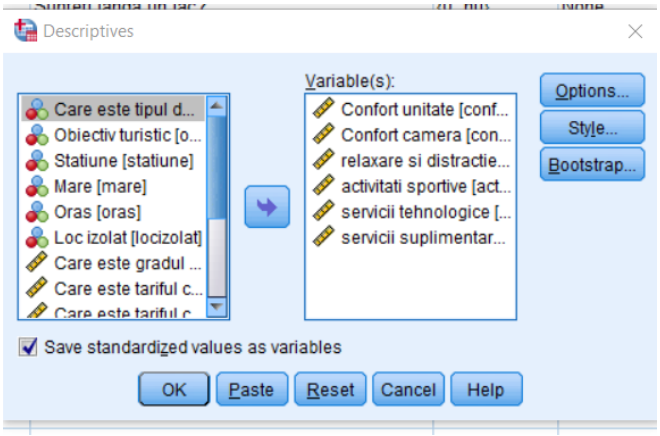
În felul acesta vedem cine este această componentă, mai exact combinații liniare de itemi inițiali.

Analiza K-Means

Analiza K-means este o metodă care intenționează să unească n variabile în k clustere, unde fiecare variabilă aparține clusterului cu media cea mai apropiată. Obiectivul principal este de a minimiza variația din interiorul clusterelor, obținând astfel grupe omogene. Mai exact, grupele să fie omogene la interior și eterogene la exterior.

Pentru realizarea acestei metode am folosit variabilele ”confort unitate”, ”confort cameră”, ”relaxare și distracții”, ”activități sportive”, ”servicii tehnologice”, ”servicii suplimentare” din baza de date ”baza 3.csv” (primită la examenul de la statistică inferențială). Aceste variabile ar putea fi de folos, să zicem, la un studiu asupra alegerii unei anumite stațiuni sau tip de cazare.

Am rulat analiza atât pe variabilele nestandardizate, cât și pe cele standardizate și observat că analiza e mai coerentă din punctul meu de vedere pe datele standardizare, așadar am decis să continui în prealabil. Pentru standardizare am mers la: Analyze – Descriptive Statistics – Descriptives, am trecut variabilele în căsuța ”Variables” și am bifat opțiunea ”Save standardized values as variables”.



Astfel, s-au creat variabilele standardizate salvate cu denumirea lor la care se adaugă Zscore.

Calea algoritmului este Analyze – Classify – K-means cluster. Introducem toate cele 6 variabile în căsuța ”Variables” și aplicăm următoarele setări:

- Number of Clusters (numărul clusterelor): 4
- Iterate: 20 (maximul este de 999)
- Save: Cluster membership, Distance from cluster center

- Options: Initial cluster centers, ANOVA table, Cluster information for each case

Analiza începe cu tabelul ”Initial Cluster Centers” și arată centrul inițial al clusterelor

Initial Cluster Centers

	Cluster			
	1	2	3	4
Zscore: Confort unitate	-1.53448	-3.13092	1.65842	.59412
Zscore: Confort camera	2.10596	-.94689	2.10596	-.94689
Zscore: relaxare si distractie	-1.67626	-1.09887	2.36551	1.21072
Zscore: activitati sportive	-.67032	-.67032	1.63510	2.78781
Zscore: servicii tehnologice	2.28684	-1.63864	2.28684	-1.63864
Zscore: servicii suplimentare	1.33821	-.99760	2.89541	.55961

Pe baza datelor am putea spune că rata cea mai ridicată a confortului unității se află în clusterul 3, iar pe de altă partea rata cea mai scăzută în clusterul 2. Se poate observa că, în general, clusterul 3 cuprinde cele mai ridicate rate ale variabilelor, iar clusterul 2 cuprinde cele mai scăzute rate, clusterele 1 si 4 cuprind valorile medii.

Tabelul ”Iteration History” indică istoricul iterațiilor

Iteration History[a]

	Change in Cluster Centers			
Iteration	1	2	3	4
1	2.860	2.712	2.276	2.809
2	.117	.083	.459	.174
3	.068	.041	.191	.119
4	.043	.049	.099	.075
5	.025	.022	.038	.038
6	.019	.009	.000	.016
7	.000	.022	.000	.020
8	.000	.013	.000	.011
9	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 9. The minimum distance between initial centers is 5.750.

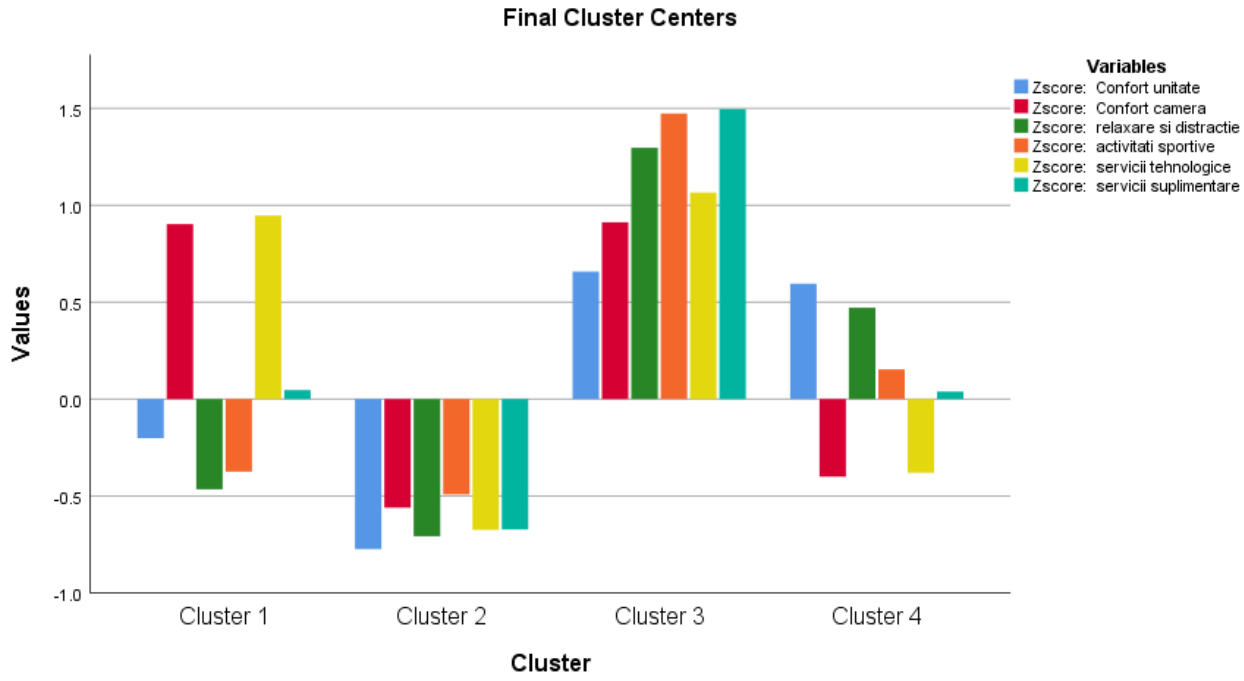
Istoricul iterațiilor este algoritmul propriu-zis al metodei. După cum se poate observa, algoritmul s-a oprit la iterația 9, deoarece în toate cele 4 clustere s-a atins valoarea 0.000, însemnând că s-au stabilizat grupele.

Tabelul ”Final Cluster Centers” arată centrul final al clusterelor

Final Cluster Centers

	Cluster			
	1	2	3	4
Zscore: Confort unitate	-.20187	-.77404	.65719	.59555
Zscore: Confort camera	.90278	-.56032	.91196	-.40051
Zscore: relaxare si distractie	-.46567	-.70733	1.29626	.47146
Zscore: activitati sportive	-.37488	-.49230	1.47287	.15305
Zscore: servicii tehnologice	.94673	-.67469	1.06558	-.38041
Zscore: servicii suplimentare	.04599	-.67280	1.49585	.03914

Final Cluster Centers ne arată centrul fiecărei variabile în cele 4 clustere dupa ce s-au produs toate modificarile și convergența a fost atinsă. Putem afirma că ratele cele mai mare se înregistrează tot în clusterul 3, iar ratele în mare măsură cele mai scăzute în clusterul 2.



Graficul de tip bar este redat pentru o mai bună vizualizare a clusterelor.

Testul ANOVA:

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore: Confort unitate	134.034	3	.630	1078	212.827	.000
Zscore: Confort camera	157.291	3	.565	1078	278.366	.000
Zscore: relaxare si distractie	176.673	3	.511	1078	345.664	.000
Zscore: activitati sportive	138.972	3	.616	1078	225.591	.000
Zscore: servicii tehnologice	191.385	3	.470	1078	407.052	.000
Zscore: servicii suplimentare	152.047	3	.580	1078	262.310	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Testul ANOVA este folosit pentru validarea clusterizării.

- Se formulează ipoteza nulă H_0: Rata confortului unității nu se diferențiază la nivelul celor 4 clustere.
- Se formulează ipoteza alternativă H_1: Cel puțin 2 valori se diferențiază semnificativ la nivelul celor 4 clustere.
- ”Mean Square” din coloana ”Cluster” semnifică varianța dintre clustere, iar ”Mean Square” din coloana ”Error” semnifică varianța din clustere. F este calculat ca și raport între cele două. Cu cât F este mai mare, cu atât variabila este mai importantă în cadrul analizei.
- ”Sig” este probabilitatea de acceptare a ipotezei nule atunci când aceasta e adevărată. Cu cât Sig e mai mare, cu atât variabila este mai neimportantă. În cazul nostru: Sig = 0.000 < pragul de semnificație de 5% de unde rezultă că ipoteza nulă se respinge, iar ipoteza alternativă se acceptă, cel puțin 2 valori se diferențiază semnificativ la nivelul celor 4 clustere.

Asemănător se face și în cazul celorlalte cinci variabile.

Numărul cazurilor în fiecare cluster:

Number of Cases in each Cluster

Cluster	1	238.000
	2	338.000
	3	135.000
	4	371.000
Valid		1082.000
Missing		.000

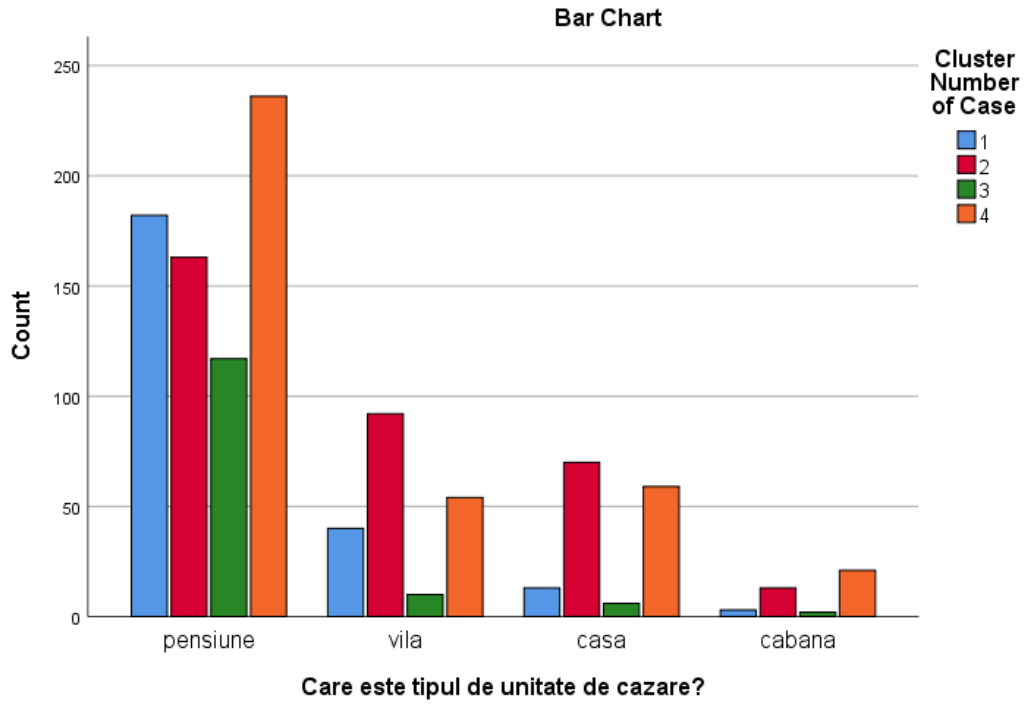
- Primul cluster conține 238 de cazuri;
- Al doilea cluster conține 338 de cazuri;
- Al treilea cluster conține 135 de cazuri și este cel mai mic;
- Al patrulea cluster conține 371 de cazuri și este cel mai voluminos.

Pentru explicarea clusterelor am făcut un tabel încrucișat între ”care este tipul de unitate de cazare” și variabila nou creată ”cluster number of case”

Care este tipul de unitate de cazare? * Cluster Number of Case Crosstabulation

			Cluster Number of Case				Total
			1	2	3	4	
Care este tipul de unitate de cazare?	pensiune	Count	182	163	117	236	698
		% within Cluster Number of Case	76.5%	48.2%	86.7%	63.8%	64.6%
	vila	Count	40	92	10	54	196
		% within Cluster Number of Case	16.8%	27.2%	7.4%	14.6%	18.1%
	casa	Count	13	70	6	59	148
		% within Cluster Number of Case	5.5%	20.7%	4.4%	15.9%	13.7%
	cabana	Count	3	13	2	21	39
		% within Cluster Number of Case	1.3%	3.8%	1.5%	5.7%	3.6%
Total		Count	238	338	135	370	1081
		% within Cluster Number of Case	100.0%	100.0%	100.0%	100.0%	100.0%

Și am bifat și graficul pentru a vizualiza mai ușor datele.



După cum se poate observa, unitatea de cazare de tip ”pensiune” deține cele mai multe înregistrări, în timp ce unitatea de tip ”cabană” deține cele mai puține înregistrări, iar unitățile de cazare ”vila” și ”casa” se aseamănă atât din punct de vedere al înregistrărilor, cât și din punct de vedere al cazurilor per cluster. Pentru pensiune avem 182 de cazuri în clusterul 1, 163 de cazuri în clusterul 2, 117 cazuri în clusterul 3 și 236 cazuri în clusterul 4. Vedem că cele mai multe cazuri sunt în clusterul 4 de unde deducem că persoanele aleg pensiunile pentru confortul unității, relaxare și distracție și pentru activități sportive. Mai putem spune că în clusterul 1 avem într-o proporție de 76.5% unități de cazare de tip pensiuni și într-o proporție de doar 1.3% unități de cazare de tip cabane.

Clusterizarea ierarhică

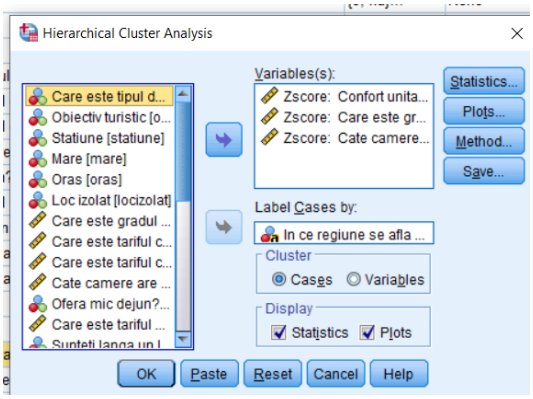
Prin analiza cluster se urmărește, în interiorul unor mulțimi de obiecte sau forme, identificarea de clase, grupe sau clustere cu elementele cât mai asemănătoare în interiorul aceleiași clase (variabilitate minimă în interiorul claselor) și cât mai diferite între ele dacă aceste elemente aparțin unor clase diferite (variabilitate maximă între clase).

Clusterizarea ierarhică pornește fără existența unor informații cu privire la numărul de clase și la apartenența formelor la aceste clase. În acest caz, clasele se construiesc pe măsura creșterii numărului de forme analizate, numărul de clase posibile determinându-se la finalul procesului de recunoaștere.

Pentru această metodă am folosit tot ”baza 3.sav” și variabilele ”câte camere are pensiunea?”, ”confort unitate”, ”gradul de mulțumire al turiștilor?”. Deoarece variabilele au scale diferite, le-am standardizat pentru a acorda aceeași importanță fiecăreia. Pentru standardizare am mers la: Analyze – Descriptive Statistics – Descriptives, am trecut variabilele în căsuța ”Variables” și am bifat opțiunea ”Save standardized values as variables”.

Calea metodei: Analyze – Classify – Hierarchical Cluster. Acolo facem următoarele modificări:

- Trecem cele trei variabile la ”Variables”
- Pentru label trecem variabila ”în ce regiune se află pensiunea?”
- Plots: Dendogram
- Method: Ward s method
- Measure: Squared Euclidian distance



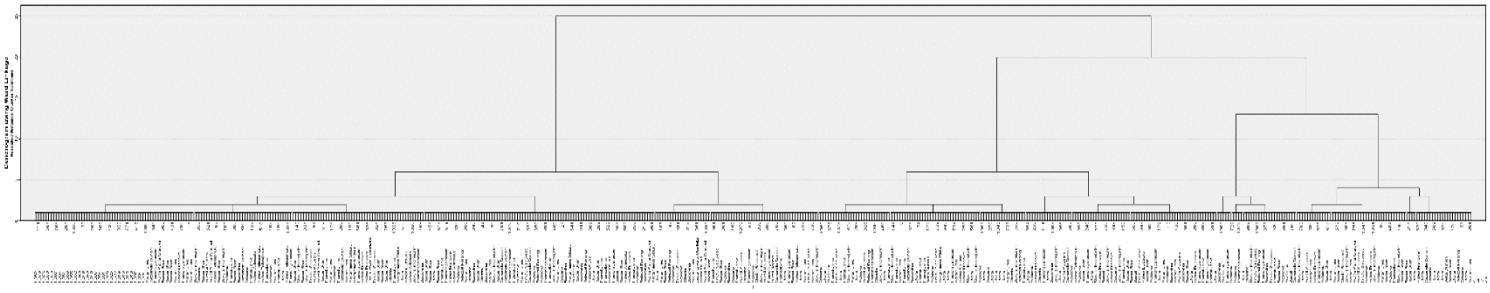
Tabelul ”Agglomeration Schedule”

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	821	1077	.000	0	0	272
2	649	1073	.000	0	0	723
3	866	1064	.000	0	0	214
4	741	1028	.000	0	0	140
5	938	1012	.000	0	0	179
6	46	1004	.000	0	0	141
7	942	971	.000	0	0	364
8	291	964	.000	0	0	325
9	242	955	.000	0	0	425
10	820	947	.000	0	0	395

După cum se poate observa cea mai mică distanța a fost între înregistrarea 821 și 1077, astfel s-a creat prima grupă, iar cele două se vor grupa în continuare la poziția 272. Următoarea combinație a fost făcută între unitatea 469 și 1073, iar cele două se vor grupa în continuare la poziția 723.

Numărul înregistrările este foarte mare, de aceea punerea graficului ”vertical icicle” aici este imposibilă.



Dendograma

Dendograma a fost pusă la o dimensiune foarte mică pentru a încăpea în document și a se observa clusterele formate. Pe baza datelor se pot forma maxim 18 clustere, însă nu este deloc recomandat. Vom alege să rămânem cu un număr de 6 clustere mai mari și să repetăm analiza. De data aceasta alegem:

- Single Solution = 6
- Save

Așa s-a creat o nouă variabilă ”Clu6_1” care memorează apartenența fiecărei unități la unul din cele 6 clustere. Mai departe vom ieși din zona de clusterizare ierarhică și vom face un tabel de frecvență pentru a vedea câte înregistrări are fiecare cluster. Analyze – Descriptive Statistics – Frequencies, iar la ”Variables” vom trece noua variabilă.

Ward Method

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	132	12.2	12.2	12.2
	2	467	43.2	43.2	55.4
	3	144	13.3	13.3	68.7
	4	156	14.4	14.4	83.1
	5	60	5.5	5.5	88.6
	6	123	11.4	11.4	100.0
	Total	1082	100.0	100.0	

După cum se poate observa, în primul cluster sunt 132 de înregistrări din totalul de 1082 sau mai putem spune ca în primul cluster sunt 12.2% din totalul înregistrărilor. Al doilea cluster este cel mai cuprinzător și deține 467 de cazuri sau 43.2% din totalul cazurilor. Numărul de cazuri diferă de la cluster la cluster pentru că urmăream împărțirea datelor în funcție de variabilele de interes, nu formarea egală a clusterelor.

Pentru validarea datelor facem analiza ANOVA. La ”Dependent List” trecem variabilele standardizate utilizate pentru clusterizare, iar la ”Factor” trecem noua variabilă, adică apartenența la cluster.

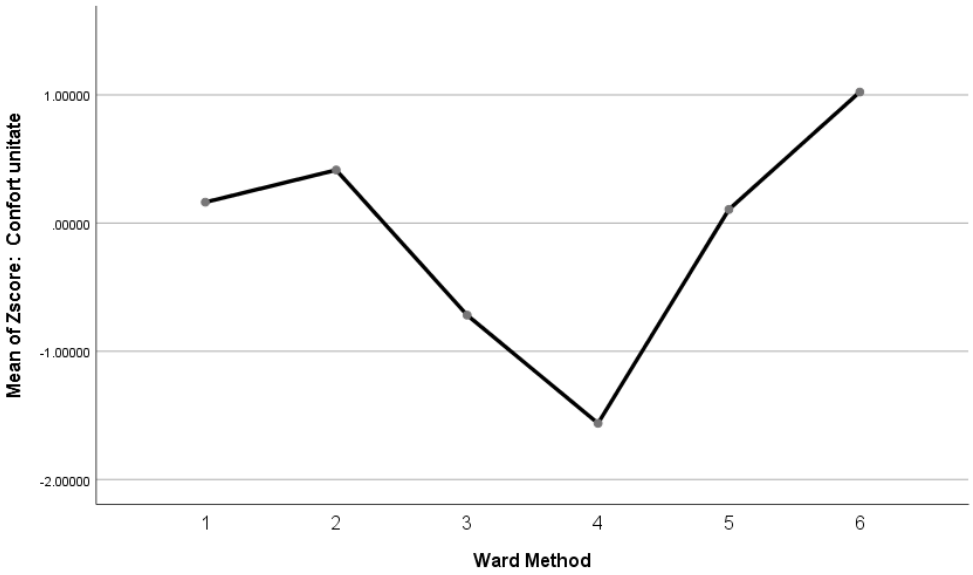
ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Zscore: Confort unitate	Between Groups	667.518	5	133.504	347.415	.000
	Within Groups	413.482	1076	.384		
	Total	1081.000	1081			
Zscore: Care este gradul de multumire al turistilor?	Between Groups	671.406	5	134.281	352.756	.000
	Within Groups	409.594	1076	.381		
	Total	1081.000	1081			
Zscore: Cate camere are pensiunea?	Between Groups	669.110	5	133.822	349.589	.000
	Within Groups	411.890	1076	.383		
	Total	1081.000	1081			

- Se formulează ipoteza nulă H_0: media variabilelor nu difera semnificativ la nivelul celor 6 clustere.

- Se formulează ipoteza alternativă H_1: Cel puțin 2 valori se diferențiază semnificativ la nivelul celor 6 clustere.
- ”Mean Square” din coloana ”Cluster” semnifică varianța dintre clustere, iar ”Mean Square” din coloana ”Error” semnifică varianța din clustere. F este calculat ca și raport între cele două. Cu cât F este mai mare, cu atât variabila este mai importantă în cadrul analizei.
- ”Sig” este probabilitatea de acceptare a ipotezei nule atunci când aceasta e adevărată. Cu cât Sig e mai mare, cu atât variabila este mai neimportantă. În cazul nostru: Sig = 0.000 < pragul de semnificație de 5% de unde rezultă că ipoteza nulă se respinge pentru toate variabilele, iar ipoteza alternativă se acceptă, cel puțin 2 valori se diferențiază semnificativ la nivelul celor 4 clustere.

Aceste test ne spune că metoda de clusterizare ierarhică poate fi validată pentru toate variabilele folosite.



Din acest grafic pentru interpreta faptul că în clusterul 4 se găsesc unitățile de cazare cu cel mai mic grad de confort, iar în clusterul 6 se găsesc unitățile de cazare cu cel mai mare grad de confort.

Cu toate acestea, nu e suficient să zicem unde se găsesc cel mai mare și cel mai mic grad, ci ne interesează valorile exacte. Pentru acest lucru mergem la Analyze – Compare means – Means,

iar aici la ”Dependent List” trecem variabilele nestandardizate și ca valoare independent va fi numărul clusterului.

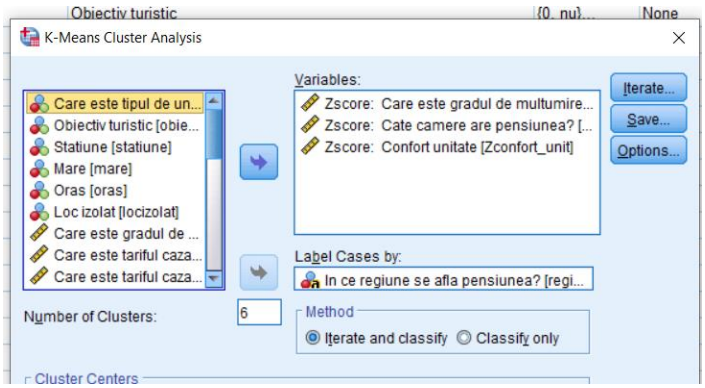
Report

Ward Method		Cate camere are pensiunea?	Confort unitate	Care este gradul de multumire al turistilor?
1	Mean	8.53	.71894	6.9323
	N	132	132	132
	Std. Deviation	3.249	.184983	1.05879
2	Mean	6.82	.76617	9.3153
	N	467	467	467
	Std. Deviation	2.286	.086236	.45043
3	Mean	13.44	.55347	8.7006
	N	144	144	144
	Std. Deviation	4.199	.137873	.58916
4	Mean	7.72	.39487	9.2833
	N	156	156	156
	Std. Deviation	3.552	.111178	.41451
5	Mean	24.18	.70833	7.9280
	N	60	60	60
	Std. Deviation	7.060	.161866	1.12804
6	Mean	12.79	.88049	9.1625
	N	123	123	123
	Std. Deviation	3.320	.062277	.43708
Total	Mean	9.68	.68835	8.8439
	N	1082	1082	1082
	Std. Deviation	5.527	.187917	1.00785

Din acest reportu putem astfel interpreta ca în clusterul 1 sunt unitățile de cazare cu aproximativ 9 camere, confortul uitații de 0.71 și gradul de mulțumire al turiștilor 6.93. În clusterul 2 se regăsesc unitățile de cazare cu 7 camere, confortul de 0.76 și graful de mulțumire al turiștilor 9.31 și așa mai departe. De altfel am putea spune ca în clusterul 5 se află un segment mai eterogen în raport cu numărul de camere, deviația standard fiind cea mai mare, 7.06. În clusterul 1 se află un segment mai eterogen în raportu cu confortul unității de cazare, deviația standard fiind cea mai

mare, 0.18. Tot în clusterul 5 mai avem și eterogenitate în raport cu gradul de mulțumire al turiștilor, deviația standard fiind 1.12, cea mai mare.

Pentru continuarea validării am făcut și analiza k-means. În analiză am inclus variabilele folosite la clusterizarea ierarhică, însă am fixat numărul exact de clustere, adică 6.



Voi interpreta într-o manieră mai sistematică pentru a nu mă repeta cu explicațiile de la pasul anterior când am făcut din nou analiza k-means.

Iteration History[a]

Change in Cluster Centers

Iteration	1	2	3	4	5	6
1	1.802	1.581	1.512	1.711	1.784	1.843
2	.387	.307	.080	.672	.447	.385
3	.110	.113	.038	.271	.203	.200
4	.056	.045	.035	.000	.191	.178
5	.036	.061	.017	.184	.220	.105
6	.028	.066	.012	.000	.151	.056
7	.039	.037	.010	.000	.217	.024
8	.038	.023	.015	.000	.135	.041
9	.070	.011	.022	.251	.358	.034
10	.057	.011	.014	.211	.325	.077
11	.029	.000	.005	.099	.268	.042
12	.038	.014	.005	.086	.262	.096
13	.012	.010	.003	.249	.194	.040
14	.038	.007	.003	.140	.235	.024
15	.036	.010	.005	.063	.217	.026
16	.032	.023	.000	.077	.158	.013
17	.024	.008	.004	.000	.080	.019

18	.022	.012	.000	.119	.158	.023
19	.019	.000	.000	.000	.052	.000
20	.030	.000	.000	.000	.105	.017
21	.019	.000	.000	.000	.047	.000
22	.006	.000	.000	.000	.015	.000
23	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 23. The minimum distance between initial centers is 3.974.

Clusterele s-au stabilizat în cea de-a 23-a iterație, valorile indicând 0.000 pentru toate clusterele.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore: Care este gradul de multumire al turistilor?	158.683	5	.267	1076	593.718	.000
Zscore: Cate camere are pensiunea?	116.228	5	.465	1076	250.191	.000
Zscore: Confort unitate	145.863	5	.327	1076	446.279	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Testul ANOVA este folosit pentru validarea clusterizării.

H_0:„ mediile de la nivelul clusterelor nu se diferențiază semnificativ”.

Testul F are valori destul de mari pentru toate variabilele și Sig = 0.000 < 0.05 (pragul de semnificație), de unde rezultă că ipoteza nulă se respinge, iar mediile valorilor dintre clustere se diferențiază. Putem spune, deci, că avem o analiză de calitatate.