Proiect final Analiza Datelor

Andreea-Adriana Suciu
1-17-2022

Cuprins

Analiza corespondențelor simplă

Pentru a rula această analiză am folosit o bază de date rezultată în urma unui studiu despre factorii care au impact asupra psihicului respondenților și a stării de bine. Chestionarul conține întrebări legate de cauzele stresului, metodele prin care fac față stresului, influența stresului în activitățile de zi cu zi. De asemenea, s-au măsurat scale legate de optimism, stimă de sine, percepția controlului, efecte pozitive și negative și satisfacția asupra stilului de viață.

Chestionarul a fost împărțit publicului general dintr-un oraș din Australia. S-au strâns 439 de răspunsuri, dintre care 42% au fost date de bărbați, 58% de femei, cu vârste cuprinse între 18 și 82 de ani (media acestora fiind de 37.4 ani).

În continuare voi prezenta analiza efectuată asupra variabilelor cauza stresului (variabilă nominală) și ultima formă de învățământ absolvită (variabilă ordinală).

→ În primul rând, ne propunem să analizăm dacă există asociere între cele 2 variabile.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
highest educ completed * source of stress	422	96,1%	17	3,9%	439	100,0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	81,960[a]	40	,000
Likelihood Ratio	75,625	40	,001
Linear-by-Linear Association	8,003	1	,005
N of Valid Cases	422		

a. 33 cells (61,1%) have expected count less than 5. The minimum expected count is ,06.

Formulăm ipoteza nulă H_0 = Nu există asociere între cele 2 variabile.

În urma calculelor făcute, constatăm că $\chi^2 = 81.960$ cu o probabilitate de acceptare a ipotezei nule când ea este adevărată, p-value = 0.000 și 40 de grade de libertate. La un prag de 5% constatăm că ipoteza nulă se respinge, în consecință există asociere între cele 2 ultima formă de învățământ absolvită și cauza stresului.

→ Avem tabelul de contingență ce reflectă distribuția respondenților în raport cu cele două variabile analizate.

Correspondence Table

	source of stress									
highest educ completed	WORK	SPOUSE OR PARTNER	RELATIONSHIPS	CHILDREN	FAMILY	HEALTH/ILLNESS	LIFE IN GENERAL	MONEY/FINANCES	TIME (lack of time, too much to do)	Active Margin
PRIMARY	0	0	0	1	0	1	0	0	0	2
SOME SECONDARY	12	5	0	6	7	5	5	7	1	48
COMPLETED HIGHSCHOOL	45	1	3	4	6	4	7	10	5	85
SOME ADDITIONAL TRAINING	56	2	4	6	8	5	8	23	1	113
COMPLETED UNDERGRADUATE	80	2	5	5	5	1	10	7	5	120
POSTGRADUATE COMPLETED	31	2	0	3	1	4	2	7	4	54
Active Margin	224	12	12	25	27	20	32	54	16	422

Având ca punct de pornire aceste date se vor calcula frecvenţele condiţionate pentru fiecare variabilă în parte.

→ Astfel avem aceste frecvențe pentru prima variabilă, denumite și „profilele” liniilor.

Row Profiles

	source of stress									
highest educ completed	WORK	SPOUSE OR PARTNER	RELATIONSHIPS	CHILDREN	FAMILY	HEALTH/ILLNESS	LIFE IN GENERAL	MONEY/FINANCES	TIME (lack of time, too much to do)	Active Margin
PRIMARY	,000	,000	,000	,500	,000	,500	,000	,000	,000	1,000
SOME SECONDARY	,250	,104	,000	,125	,146	,104	,104	,146	,021	1,000
COMPLETED HIGHSCHOOL	,529	,012	,035	,047	,071	,047	,082	,118	,059	1,000
SOME ADDITIONAL TRAINING	,496	,018	,035	,053	,071	,044	,071	,204	,009	1,000
COMPLETED UNDERGRADUATE	,667	,017	,042	,042	,042	,008	,083	,058	,042	1,000
POSTGRADUATE COMPLETED	,574	,037	,000	,056	,019	,074	,037	,130	,074	1,000
Mass	,531	,028	,028	,059	,064	,047	,076	,128	,038	

Din acest tabel observăm că respondenții care au absolvit ciclul gimnazial au ca motive ale stresului locul de muncă, copiii, familia, starea de sănătate, viața în general și gestionarea banilor; cei care au absolvit învățământul preuniversitar (completed undergraduate) sunt stresați, în mare parte, din cauza locului de muncă.

Sau putem observa că 52.9% ($\frac{45}{85} * 100\%$) dintre cei care au absolvit liceul recunosc că sursa principală care le cauzează stres este locul de muncă.

→ În continuare analizăm frecvențele condiționale pentru variabila „sursa stresului”, adică „profilele” coloanelor.

Column Profiles

	source of stress									
highest educ completed	WORK	SPOUSE OR PARTNER	RELATIONSHIPS	CHILDREN	FAMILY	HEALTH/ILLNESS	LIFE IN GENERAL	MONEY/FINANCES	TIME (lack of time, too much to do)	Mass
PRIMARY	,000	,000	,000	,040	,000	,050	,000	,000	,000	,005
SOME SECONDARY	,054	,417	,000	,240	,259	,250	,156	,130	,063	,114
COMPLETED HIGHSCHOOL	,201	,083	,250	,160	,222	,200	,219	,185	,313	,201
SOME ADDITIONAL TRAINING	,250	,167	,333	,240	,296	,250	,250	,426	,063	,268
COMPLETED UNDERGRADUATE	,357	,167	,417	,200	,185	,050	,313	,130	,313	,284
POSTGRADUATE COMPLETED	,138	,167	,000	,120	,037	,200	,063	,130	,250	,128
Active Margin	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	

De aici rezultă: persoanele care sunt stresate din cauza muncii, relațiilor, timpului, vieții în general au absolvit învățământul preuniversitar. Iar cei care sunt stresați din cauza banilor și relațiilor au urmat cursuri adiționale.

Sau putem spune că 41.7% ($\frac{5}{12} * 100\%$) dintre cei care sunt stresați din cauza relațiilor au absolvit studii preuniversitare.

→ În următorul tabel ne sunt prezentate numărul maxim admis al dimensiunilor în care se împarte informația și proporția acesteia în fiecare dimensiune/axă.

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	,340	,116			,595	,595	,049	,197
2	,187	,035			,180	,774	,046	
3	,154	,024			,123	,897		
4	,121	,015			,076	,973		
5	,073	,005			,027	1,000		
Total		,194	81,960	,000[a]	1,000	1,000		

a. 40 degrees of freedom

Din acest tabel putem spune că 59.5% ($\frac{0.116}{0.194} * 100\%$) din informația totală se va regăsi în dimensiunea 1, 18.0% în dimensiunea 2 ș.a.m.d. În același timp putem interpreta frecvențele cumulate, adică 89.7% din informația inițială se va regăsi în primele 3 dimensiuni/axe.

→ Interpretarea rezultatelor și aprecierea calității lor se realizează prin intermediul contribuțiilor liniilor („Overview Row Points”) și coloanelor („Overview Column Points”) la varianța dimensiunilor/axelor, regăsite în tabelele de mai jos.

Overview Row Points[a]

highest educ completed	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
PRIMARY	,005	3,453	-3,748	,040	,166	,356	,477	,309	,786
SOME SECONDARY	,114	1,303	,181	,073	,568	,020	,895	,009	,904
COMPLETED HIGHSCHOOL	,201	-,132	-,067	,006	,010	,005	,217	,031	,247
SOME ADDITIONAL TRAINING	,268	,065	,489	,021	,003	,343	,019	,580	,599
COMPLETED UNDERGRADUATE	,284	-,548	-,157	,037	,251	,038	,786	,035	,821
POSTGRADUATE COMPLETED	,128	,004	-,590	,017	,000	,239	,000	,477	,477
Active Total	1,000			,194	1,000	1,000			

a. Symmetrical normalization

Overview Column Points[a]

source of stress	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
WORK	,531	-,399	-,104	,030	,249	,031	,947	,035	,982
SPOUSE OR PARTNER	,028	1,330	,143	,028	,148	,003	,614	,004	,618
RELATIONSHIPS	,028	-,706	,432	,010	,042	,028	,504	,104	,608
CHILDREN	,059	,989	-,547	,026	,171	,095	,757	,127	,884
FAMILY	,064	,666	,674	,019	,084	,156	,511	,287	,798
HEALTH/ILLNESS	,047	1,358	-,853	,039	,257	,185	,755	,164	,919
LIFE IN GENERAL	,076	,059	,267	,005	,001	,029	,020	,223	,243
MONEY/FINANCES	,128	,299	,655	,024	,034	,294	,163	,431	,595
TIME (lack of time, too much to do)	,038	-,371	-,941	,014	,015	,180	,128	,452	,581
Active Total	1,000			,194	1,000	1,000			

a. Symmetrical normalization

Având în vedere cele 2 tabele, putem concluziona:

- Varianța dimensiunii/axei 1 se asociază cel mai bine cu munca (0.249) și starea de sănătate (0.257) drept sursă a stresului, efecte resimțite, în mare parte, de cei care au absolvit ultima oară ciclul gimnazial (0.568) și preuniversitar (0.251).
- În timp ce varianța dimensiunii/axei 2 este asociată cu stres din cauza banilor (0.294), a lipsei timpului (0.18), a stării de sănătate (0.185) și a familiei (0.156), resimțite, în mare parte, de persoanele care au terminat învățământul primar (0.356) și de cele care au urmat cursuri adiționale (0.343).

  Astfel se desprind următoarele: persoanele care au măcar studii liceale complete sunt stresați din cauza locului de muncă, în timp ce aceia care au doar o educație de bază au drept motive de stres banii, timpul, familia. Totuși, ambele categorii au o sursă comună a stresului indiferent de nivelul de educație, și anume starea de sănătate.

Analiza componentelor principale

Pentru a rula această analiză am folosit o bază de date rezultată în urma unui studiu despre factorii care au impact asupra psihicului respondenților și a stării de bine. Chestionarul conține întrebări legate de cauzele stresului, metodele prin care fac față stresului, influența stresului în activitățile de zi cu zi. De asemenea, s-au măsurat scale legate de optimism, stimă de sine, percepția controlului, efecte pozitive și negative și satisfacția asupra stilului de viață.

Chestionarul a fost împărțit publicului general dintr-un oraș din Australia. S-au strâns 439 de răspunsuri, dintre care 42% au fost date de bărbați, 58% de femei, cu vârste cuprinse între 18 și 82 de ani (media acestora fiind de 37.4 ani).

Am folosit variabila care descrie satisfacția față de viață a respondenților.

Case Processing Summary

		N	%
Cases	Valid	436	99,3
	Excluded[a]	3	,7
	Total	439	100,0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,890	,895	5

Indicatorul Cronbach's Alpha se încadrează în intervalul [0.7 ; 0.9], deci avem o situație foarte bună. În concluzie, putem aplica analiza dorită deoarece datele corespund scopului.

Item Statistics

	Mean	Std. Deviation	N
lifsat1	4,37	1,528	436
lifsat2	4,57	1,554	436
lifsat3	4,69	1,519	436
lifsat4	4,75	1,641	436
lifsat5	3,99	1,855	436

Inter-Item Correlation Matrix

	lifsat1	lifsat2	lifsat3	lifsat4	lifsat5
lifsat1	1,000	,763	,720	,573	,526
lifsat2	,763	1,000	,727	,606	,481
lifsat3	,720	,727	1,000	,721	,587
lifsat4	,573	,606	,721	1,000	,594
lifsat5	,526	,481	,587	,594	1,000

Matricea de corelație este pătratică (numărul coloanelor = numărul liniilor), pe diagonala principală avem valoarea 1 și este simetrică față de diagonala principală. Toate corelațiile liniare dintre variabile sunt pozitive. Elementele de pe diagonală, denumită urma matricii, cumulează informația inițială (5 în cazul nostru).

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
lifsat1	18,00	30,667	,758	,649	,861
lifsat2	17,81	30,496	,752	,654	,862
lifsat3	17,69	29,852	,824	,695	,847
lifsat4	17,63	29,954	,734	,574	,866
lifsat5	18,39	29,704	,627	,421	,896

Tabelul acesta ne arată statisticile descriptive și Cronbach s Alpha în caz că am renunța la una dintre variabile. Ne uităm dacă am avea o situație mai bună. Dacă vrem să avem date de mai bună calitate, atunci e bine să renunțăm la acel item. Dacă avem în tabel un coeficient Cronbach’s Alpha mai mare, dar nu cu o diferență foarte mare, este indicat să păstrăm acea variabilă.

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
22,38	45,827	6,770	5

Din cei 5 itemi vom crea o variabilă latentă.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,849
Bartlett's Test of Sphericity	Approx. Chi-Square	1332,806
	df	10
	Sig.	,000

KMO testează calitatea globală a testului.

H0: Nu există niciun fel de corelație între itemi (KMO=0)

H1: KMO diferă semnificativ de 0, ne așteptăm la o calitate bună a analizei

Probabilitatea de a accepta ipoteza nulă este 0.000 (sig), deci respingem ipoteza nulă. Acceptăm alternativa, adică ne așteptăm la o analiza de bună calitate, există corelații consistente între itemi.

KMO>0.5

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,530	70,601	70,601	3,530	70,601	70,601
2	,603	12,052	82,653			
3	,407	8,130	90,783			
4	,236	4,715	95,498			
5	,225	4,502	100,000			

Extraction Method: Principal Component Analysis.

Coloana “Total” prezintă valorile proprii calculate pornind de la matricea ”Inter-Item Correlation Matrix” și sunt ordonate descrescător în raport cu importanța lor.

Variabila latentă creată (componenta 1 din tabel) reține 70.601% ($\frac{3.530}{5} * 100$) din informația inițială. În acest fel se calculează toate valorile de pe coloana ”Initial Eigenvalues % of Variance”.

Pe următoarea coloană găsim frecvențele cumulate, de exemplu: în primele 3 componente se află 90.783% din informația totală.

Se va păstra componenta 1 pentru că reține cea mai mare parte din informația inițială, adică 70.601%.

Component Matrix[a]

	Component 1
lifsat1	,858
lifsat2	,858
lifsat3	,900
lifsat4	,831
lifsat5	,746

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

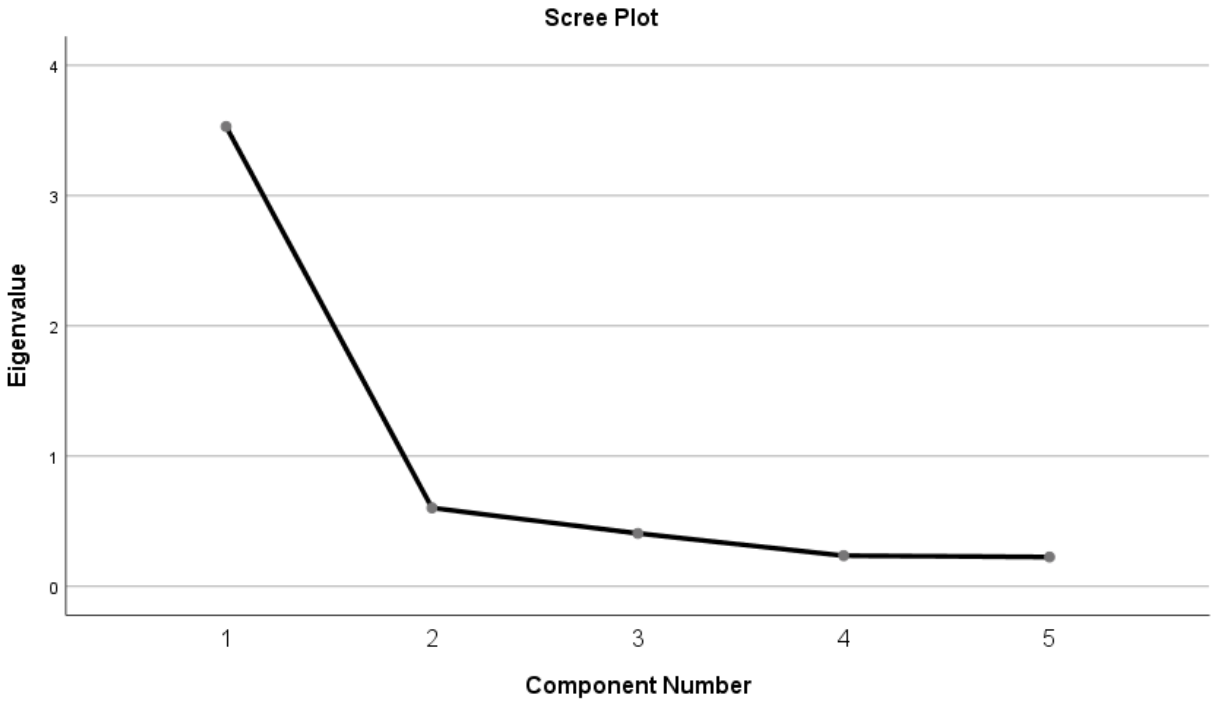
Având în vedere procentul prezentat în tabelul „Total Variance Explained” (70.601%) și la valorile coeficienților de corelație dintre fiecare item și componenta respectivă, toate fiind peste pragul de 0.5, putem concluziona că am creat o variabilă latentă de bună calitate.

Communalities

	Initial	Extraction
lifsat1	1,000	,736
lifsat2	1,000	,736
lifsat3	1,000	,810
lifsat4	1,000	,691
lifsat5	1,000	,556

Extraction Method: Principal Component Analysis.

În tabelul „Communalities” se prezintă ce procent din informația inițială se regăsește pe variabila latentă creată. Din lifsat1 s-a reținut 73.6% din informație, din lifsat3 avem 81.0% din informație ș.a.m.d.



Componenta 1 reține cea mai multă informație, iar cea mai puțină informație este reținută de componenta 5.

Component Score Coefficient Matrix

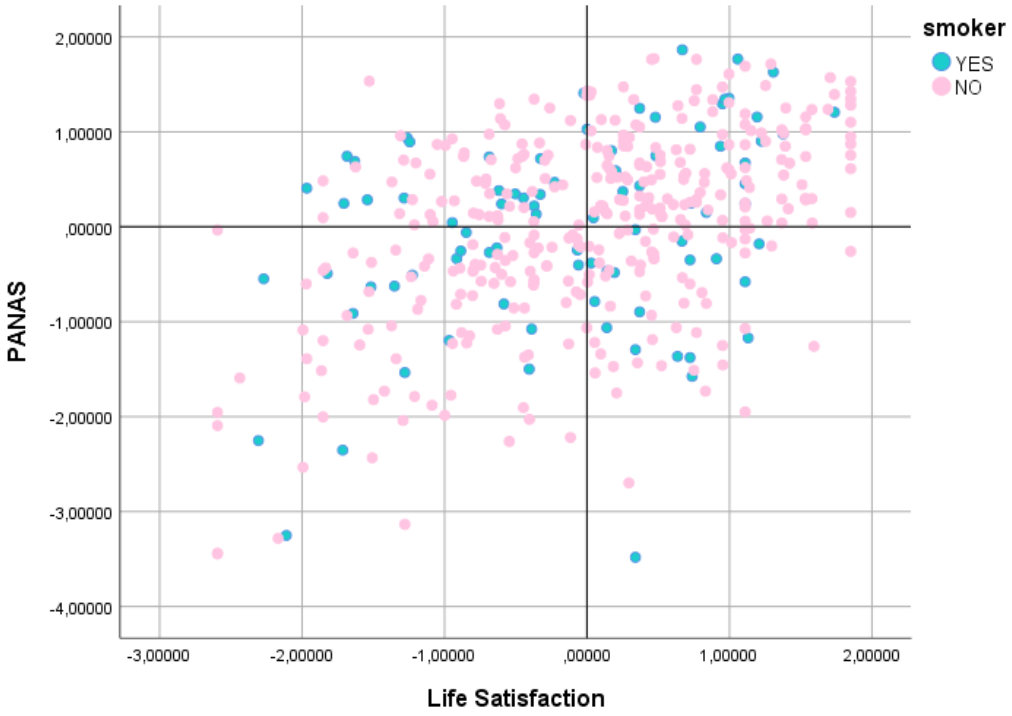
	Component 1
lifsat1	,243
lifsat2	,243
lifsat3	,255
lifsat4	,236
lifsat5	,211

Extraction Method: Principal Component Analysis.
Component Scores.

Componenta 1 = 0.243 * lifsat1 + 0.243 * lifsat2 + 0.255 * lifsat3 + 0.236 * lifsat4 + 0.211 * lifsat5

În concluzie, această variabilă latentă creată reține în memorie o cantitate mai mică de informații din total, dar este un proces eficient deoarece prezintă rezultate bune în urma testelor.

Adițional, am mai creat o variabilă latentă pentru variabila PANAS.



Conform graficului există mai multe persoane nefumătoare decât fumătoare printre respondenți. Cei mai mulți nefumători sunt mulțumiți de viața lor și au înregistrat efecte pozitive în urma testului PANAS.

K Means Cluster

Am rulat această analiză pentru a grupa respondenții chestionarului În 3 grupe în funcție de stresul perceput, satisfacția asupra vieții și stima de sine.

ANOVA						
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Total life satisfaction	6085,956	2	17,903	428	339,949	,000
Total perceived stress	3631,137	2	17,511	428	207,366	,000
Total Self esteem	3515,208	2	12,345	428	284,739	,000

Formulăm ipoteza nulă:

H_0: Satisfacția asupra vieții a respondenților nu se diferențiază pe cele 3 clustere.

H_1: Există cel puțin 2 grupe în care satisfacția medie asupra vieții se diferențiază semnificativ.

Sig=0.000 < 0.05 (pragul de semnificativitate), deci respingem ipoteza nulă, adică satisfacția medie asupra vieții se diferențiază în cel puțin 2 clustere.

Coloana Mean Square din Cluster arată varianța explicită, iar coloana Mean Square din Error arată variația reziduală.

$$F = \frac{Mean\ Square\ Cluster}{Mean\ Square\ Error} = \frac{6085{,}956}{17{,}903} = 339{,}949$$

Cu ajutorul lui F putem face clasamentul importanței variabilelor: cea mai importantă variabilă cu F=339,949 este satisfacția asupra vieții, următoarea fiind stima de sine cu F=284,739, iar cea mai puțin importantă dintre cele 3 este stresul perceput cu F=207,366.

Dacă satisfacția medie asupra vieții nu se diferențiază pe cele 3 grupe, varianța explicită va tine către zero, de unde și F-ul va tinde către zero, iar sig (probabilitatea de acceptare a ipotezei nule când este adevărată) va tinde spre 1.

Totodată, ipoteza nulă este respinsă și pentru variabilele stres perceput și stimă de sine.

Dacă avem o variabilă în urma căreia ipoteza nulă s-ar fi acceptat, adică nu se diferențiază, o putem eliminia sau o lăsăm în analiză, dar nu o interpretăm.

Number of Cases in each Cluster		
Cluster	1	187,000
	2	66,000
	3	178,000
Valid		431,000
Missing		8,000

Final Cluster Centers

	Cluster		
	1	2	3
Total life satisfaction	21	13	28
Total perceived stress	29	33	22
Total Self esteem	33	25	37

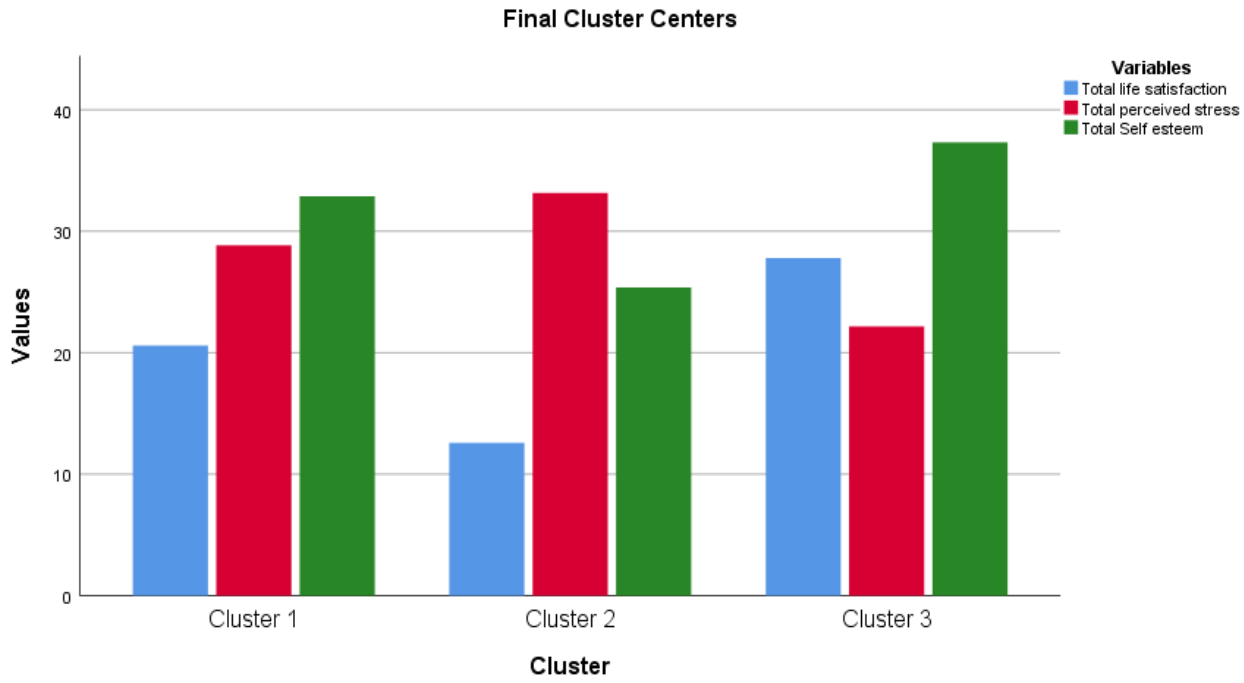
Prima grupă prezintă persoanele care simt un nivel mediu de satisfacție asupra vieții și stres. Acestea arată un nivel moderat de stimă de sine.

În a doua grupă sunt persoanele cu un grad ridicat de stres perceput care nu sunt deloc satisfăcute în legătură cu viața lor și au o stimă de sine foarte scăzută (cea mai mică dintre toate). Deci, în acest cluster avem persoane foarte stresate, fără stimă de sine și nemulțumiți de viață.

În cea de-a treia grupă avem persoanele cu o stimă de sine ridicată care sunt foarte mulțumiți și satisfăcuți de viața lor și care sunt cel mai puțin stresați dintre toate grupele.

Tabelul „Cluster Membership” ne arată fiecare observație împreună cu numărul clusterului în care a fost introdusă și, pe ultima coloană, distanța observației față de centrul clusterului respectiv.

Cu opțiunea „save cluster membership” se va salva ulterior în baza de date numărul clusterului din care face parte observația și cu opțiunea „save distance from cluster center” se va salva distanța observației față de centrul clusterului de care aparține.



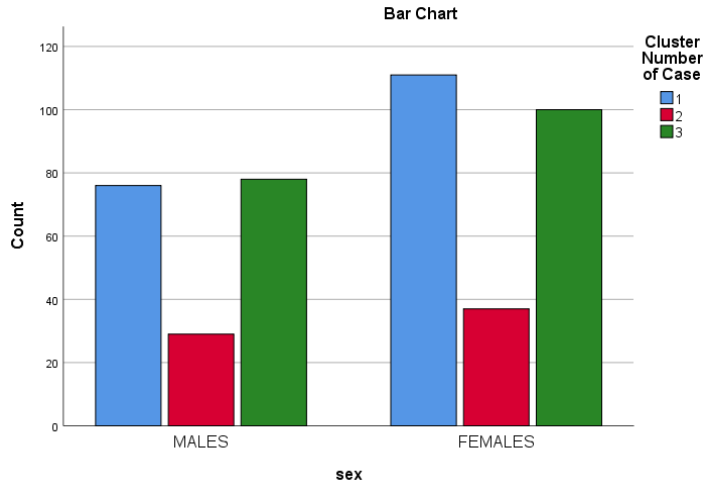
În continuare ne folosim de Crosstabs din meniul Analyze – Descriptive Statistics pentru a înțelege corelațiile între câteva variabile și variabila creată mai sus „cluster membership”.

În primul rând, analizăm sexul respondenților în legătură cu numărul clusterului din care fac parte.

După cum putem observa femeile sunt mai numeroase decât bărbații în toate clusterele create. În primul grup femeile sunt în procent de 59,4%, în cel de-al doilea 56,1% și în al treilea 56,2%.

sex * Cluster Number of Case Crosstabulation

			Cluster Number of Case			Total
			1	2	3	
sex	MALES	Count	76	29	78	183
		% within Cluster Number of Case	40,6%	43,9%	43,8%	42,5%
	FEMALES	Count	111	37	100	248
		% within Cluster Number of Case	59,4%	56,1%	56,2%	57,5%
Total		Count	187	66	178	431
		% within Cluster Number of Case	100,0%	100,0%	100,0%	100,0%



În continuare voi lua variabila „statut marital”.

marital status * Cluster Number of Case Crosstabulation

			Cluster Number of Case			Total
			1	2	3	
marital status	SINGLE	Count	53	19	31	103
		% within Cluster Number of Case	28,3%	28,8%	17,4%	23,9%
	STEADY RELATIONSHIP	Count	30	13	28	71
		% within Cluster Number of Case	16,0%	19,7%	15,7%	16,5%
	MARRIED	Count	86	23	107	216
		% within Cluster Number of Case	46,0%	34,8%	60,1%	50,1%
	DIVORCED / WIDOWED	Count	18	11	12	41
		% within Cluster Number of Case	9,6%	16,7%	6,7%	9,5%
Total		Count	187	66	178	431
		% within Cluster Number of Case	100,0%	100,0%	100,0%	100,0%

Putem spune că în primul grup domină persoanele căsătorite (46,0%), urmate de cele singure (28,3%), într-o relație stabilă (16%) și divorțate / văduve (9,6%). Și în celelalte clustere observăm același clasament.

În continuare voi rula analiza Compare Means pentru a compara grupele create între ele folosind 3 variabile: scala de optimism, scala de efect pozitiv și cea pentru efectul negativ asupra respondenților.

Report

Cluster Number of Case		Total Optimism	Total positive affect	Total negative affect
1	Mean	21,44	33,06	21,06
	Std. Deviation	3,371	5,779	6,757
2	Mean	17,48	26,50	25,12
	Std. Deviation	5,066	8,010	7,800
3	Mean	24,56	37,04	15,66
	Std. Deviation	3,513	6,191	4,780
Total	Mean	22,13	33,70	19,45
	Std. Deviation	4,448	7,257	7,096

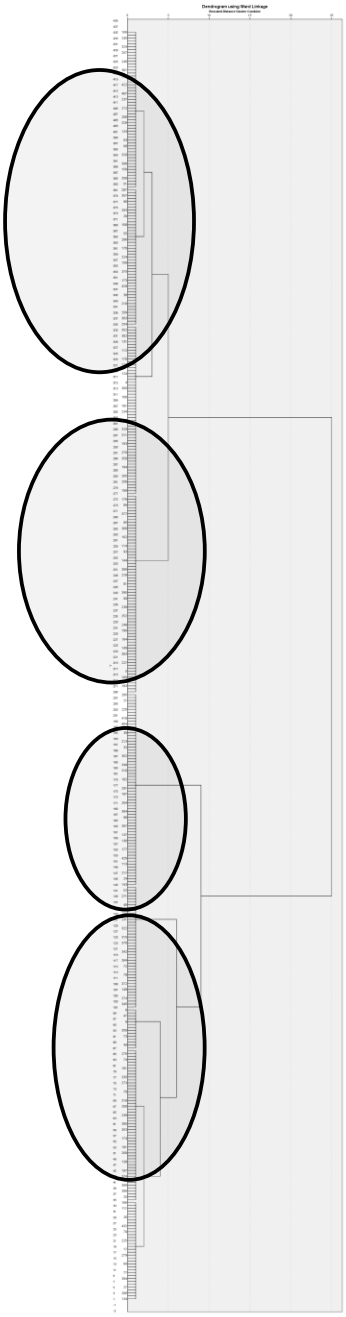
Putem observa că persoanele din primul grup creat sunt relativ optimiste și sunt afectate în mod pozitiv, cât și negativ de întâmplări într-o măsură medie. În cel de-al doilea grup avem oamenii cei mai puțin optimiști, afectați în mod negativ într-o măsură mai mare decât ceilalți. În timp ce, în al treilea grup avem cele mai optimiste persoane care sunt afectate în mod pozitiv într-o măsură mare.

În concluzie, k-means cluster împarte eșantionul în 3 grupe. Primul grup este format de persoane care simt un nivel moderat de stimă de sine, satisfacție asupra vieții și sunt ușor stresate, relativ optimiste. În cel de-al doilea grup avem persoanele foarte stresate, lipsite de stima de sine și nesatisfăcute în legătură cu modul de viață, acestea fiind cele mai puțin optimiste, afectate negativ într-o mare măsură. Iar în al treilea grup avem persoanele mulțumite de stilul lor de viață, cu stima de sine ridicată, acestea fiind cele mai optimiste și afectate pozitiv în mare parte.

În continuare voi rula atât Hierarchical Cluster, cât și K-Means Cluster pe 2 variabile create la Analiza Componentelor Principale: Life Satisfaction (analizată în document) și PANAS.

*PANAS = Planul de afecte pozitive și negative este un chestionar de auto-raport care constă din două scale de 10 elemente pentru a măsura atât efectul pozitiv cât și cel negativ. *

Prima oară vom rula Hierarchical Cluster pe cele 2 variabile. De aici ne interesează dendograma, de unde ne vom da seama câte grupe vom forma folosind procesul de clusterizare. Din dendogramă am ales să continui analizele folosind 4 clustere. În urma acestei analize am salvat „cluster membership” adică numărul clusterului în care e salvată observația.



Pentru următoarea parte vom rula analiza K-Means folosind cele 4 clustere determinate mai sus.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Life Satisfaction	104,189	3	,284	431	367,207	,000
PANAS	101,439	3	,301	431	337,128	,000

Formulăm ipoteza nulă: Satisfacția asupra vieții respondenților nu se diferențiază în cele 4 clustere.

Sig=0.000 < 0.05 => se respinge ipoteza nulă, deci în cel puțin 2 clustere satisfacția medie asupra vieții a respondenților este diferită.

Și în cazul variabilei PANAS avem aceeași situație.

Cluster	1	181,000
	2	88,000
	3	45,000
	4	121,000
Valid		435,000
Missing		4,000

În tabelul alăturat ne este prezentat numărul observațiilor din fiecare grupă creată

Avem 4 valori la „missing” deoarece algoritmul exclude observațiile fără răspuns, goale.

Primul cluster conține informație de la cei mai mulți respondenți, adică 181. Următorul, în ordinea descrescătoare a numărului de observații, este cel de-al patrulea grup cu 121 de respondenți. Pe locul al treilea avem clusterul 2 cu 88 de persoane și cele mai puține persoane sunt grupate în clusterul 3 într-un număr de 45.

marital status * Cluster Number of Case Crosstabulation

			Cluster Number of Case				Total
			1	2	3	4	
marital status	SINGLE	Count	28	27	10	39	104
		% within Cluster Number of Case	15,5%	30,7%	22,2%	32,2%	23,9%
	STEADY RELATIONSHIP	Count	29	19	9	16	73
		% within Cluster Number of Case	16,0%	21,6%	20,0%	13,2%	16,8%
	MARRIED	Count	111	39	17	50	217
		% within Cluster Number of Case	61,3%	44,3%	37,8%	41,3%	49,9%
	DIVORCED / WIDOWED	Count	13	3	9	16	41
		% within Cluster Number of Case	7,2%	3,4%	20,0%	13,2%	9,4%
Total		Count	181	88	45	121	435
		% within Cluster Number of Case	100,0%	100,0%	100,0%	100,0%	100,0%

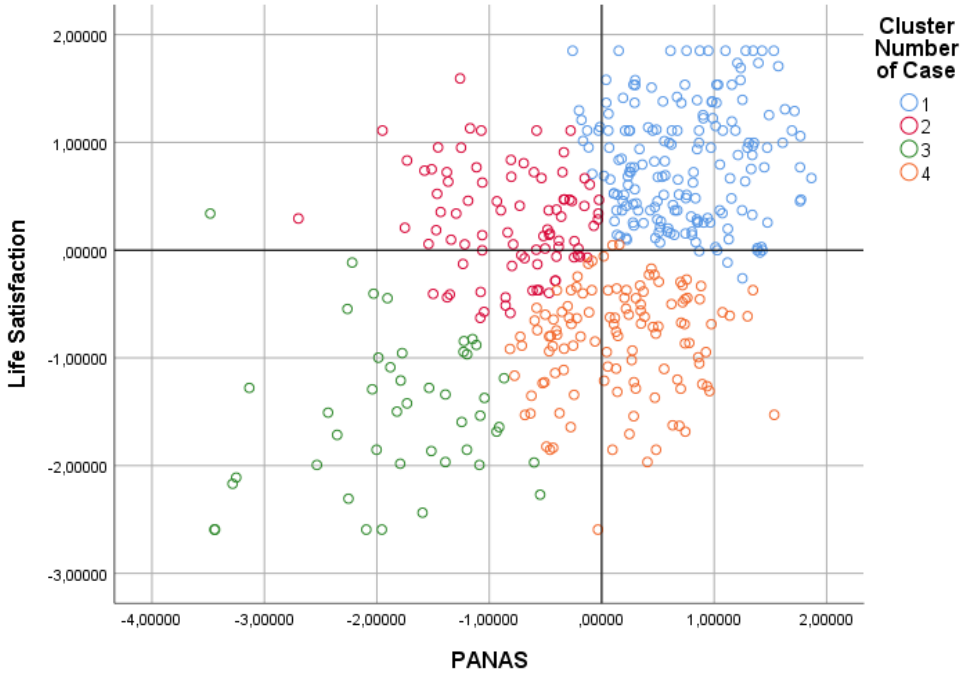
În primul cluster realizat predomină persoanele căsătorite cu un procent de 61,3% din totalul celor din această grupă. În cel de-al doilea predomină atât persoanele singure (30,7%), cât și cele căsătorite (44,3%). În a treia grupă tot procentul persoanelor căsătorite este cel mai mare, dar procentele celorlalte categorii se fluctuează în jurul valorii 20%. A patra grupă creată se aseamănă clusterului 2 cu 49,9% dintre persoane sunt căsătorite, 23,9% sunt singure, 16,8% sunt într-o relație stabilă și 9,4% sunt divorțate/văduve.

Report

age					
Ward Method	Mean	Std. Deviation	Minimum	Maximum	Median
1	38,42	13,876	18	82	36,00
2	40,39	11,952	22	67	41,00
3	34,66	11,681	18	70	33,50
4	34,07	13,366	18	69	30,00
Total	37,44	13,205	18	82	36,00

În tabelul alăturat putem vedea vârsta medie, minimă, maximă și mediană a persoanelor din fiecare cluster. Cei mai tineri îi regăsim în clusterul 4 cu o medie de 34 de ani, cel mai tânăr respondent având 18 ani și cel mai în vârsta 69. Tot în acest grup 50% dintre persoane au maxim 30 de ani, iar ceilalți au peste 30 de ani. În medie, vârstele participanților din clusterul 4 se abat de la medie cu 13,2 ani.

Grupa care are cea mai mare medie de vârstă este clusterul 2. Persoanele din această grupă se încadrează în intervalul [22;67] cu o medie de 40 de ani. Vârsta acestora se abate de la medie, în medie cu aproximativ 12 ani și jumătate dintre respondenți au maxim 41 de ani, în timp ce cealaltă jumătate au minim 41 de ani.



Conform graficului, persoanele din clusterul 1 sunt foarte asemănătoare între ele datorită poziționării punctelor atât de aproape unul de celălalt, adică grupul este omogen; aceștia având un scor mare pentru efectul pozitiv și fiind mulțumiți de viața lor. Cei din clusterul 2 sunt mulțumiți de viața lor, deși în urma testului PANAS reiese că sunt influențate negativ de ce se întâmplă în jurul lor. Cei din clusterul 3 nu se aseamănă între ei precum cei din prima grupă (adică grupul este eterogen) și aceștia sunt persoane

afectate negativ și nemulțumiți de viața lor. Iar cei din clusterul 4 au atât rezultate pozitive, cât și negative în urma testului PANAS și nu sunt satisfăcuți de stilul lor de viață.

În concluzie, respondenții au fost grupați în 4 clustere: primul format majoritar din persoane căsătorite, cu o medie de vârsta de aproximativ 38 de ani, mulțumite de viața lor și cu un scor mare pe scala de efecte pozitive. Al doilea grup are cea mai mare medie de vârstă (40 de ani) unde predomină atât persoanele căsătorite, cât și cele singure. Acestea sunt satisfăcute de stilul lor de viață, dar au tendința să fie afectate negativ de ce li se întâmplă. A treia și cea mai mică grupă este formată din respondenții nemulțumiți de viața lor, influențați negativ de întâmplările din jur. Clusterul 4 are cele mai tinere persoane care, încă, nu sunt mulțumite de viața lor, dar sunt atât afectate negativ, cât și pozitiv emoțional.

TwoStep Cluster

Pentru a aplica analiza TwoStep Cluster am folosit o baza de date rezultată în urma unui chestionar despre dependența de jocuri de noroc și impactul asupra vieții respondenților. Eșantionul este format din 134 de respondenți.

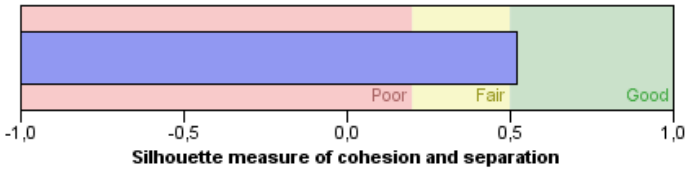
Predictorii folosiți în metodă sunt:

- Statutul social
- Cât de des obișnuiți să pariați?
- S-a întmâmplat să pierdeți și ultimul ban?
- Ați jucat la pariuri live?

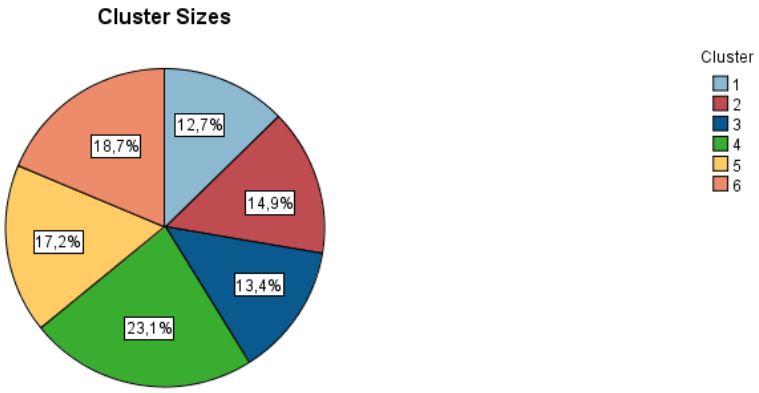
Model Summary

Algorithm	TwoStep
Inputs	4
Clusters	6

Cluster Quality

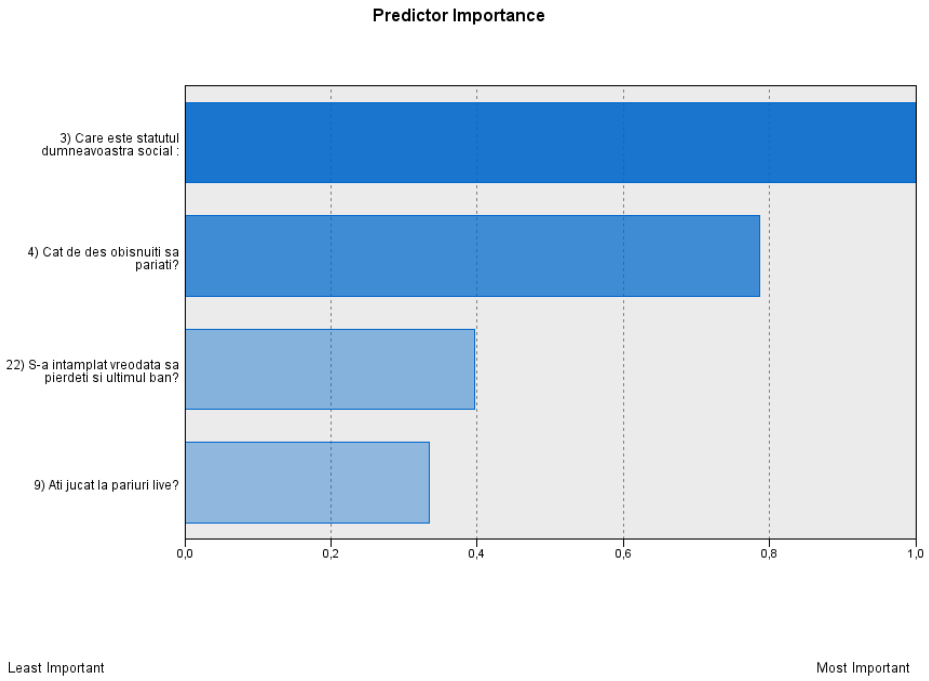


Algoritmul creează 6 clustere folosind cele 4 variabile date, iar validitatea clusterelor este măsurată prin coeficientul Silhouette care poate lua valori în intervalul [-1;1]. În cazul nostru, acest coeficient indică o bună calitate a clusterelor create.

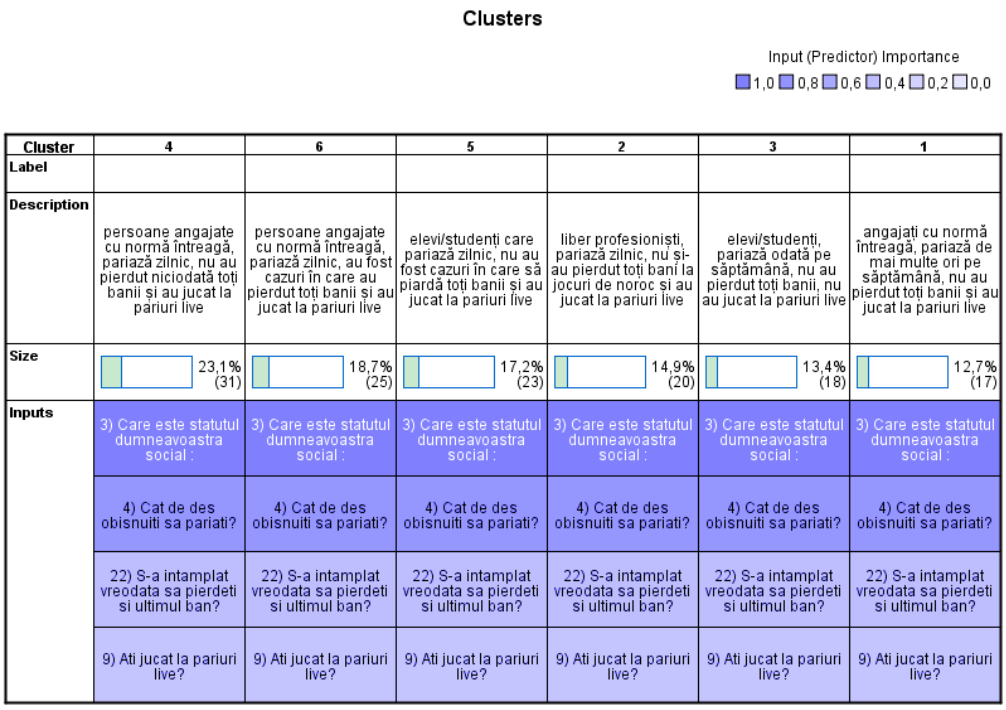


Size of Smallest Cluster	17 (12,7%)
Size of Largest Cluster	31 (23,1%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1,82

Figura anterioară ne arată dimensiunea clusterelor în totalul informației. Putem observa că cel mai mare grup conține 31 de respondenți (23,1% din total), cel mai mic este format din 17 persoane (12,7%), iar raportul de dimensiuni între cel mai mare și cel mai mic cluster este 1,82. Cele 6 clustere sunt aproximativ egale ca mărime.

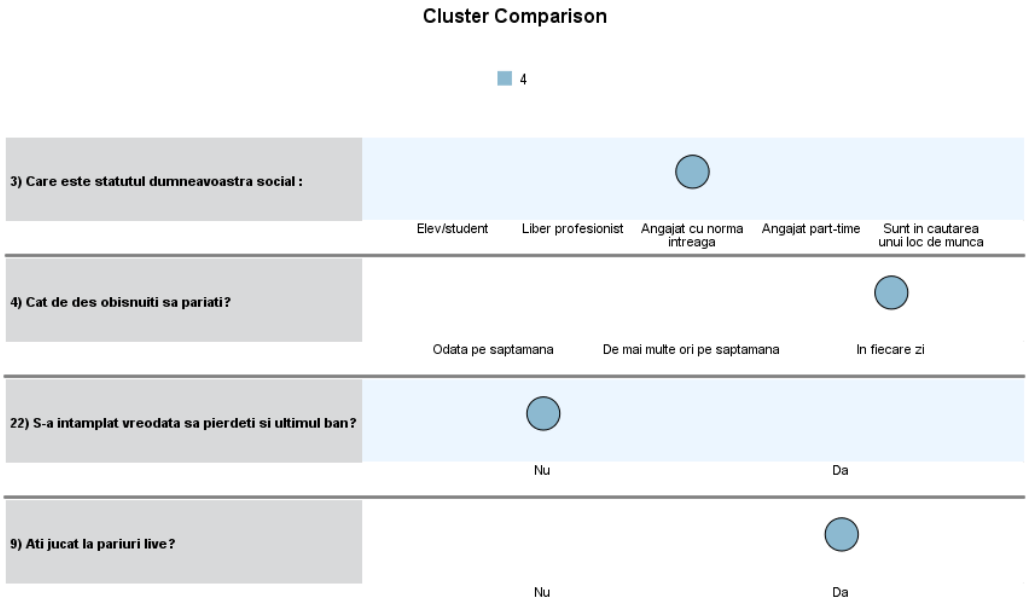


În această figură ne este prezentată importanța predictorilor pe baza cărora s-a făcut împărțirea în grupe. Cel mai important predictor este statutul social, urmat de frecvența pariurilor (0,79), posibilitatea pierderii și ultimului ban (0,40) și jocul la pariuri live (0,33).

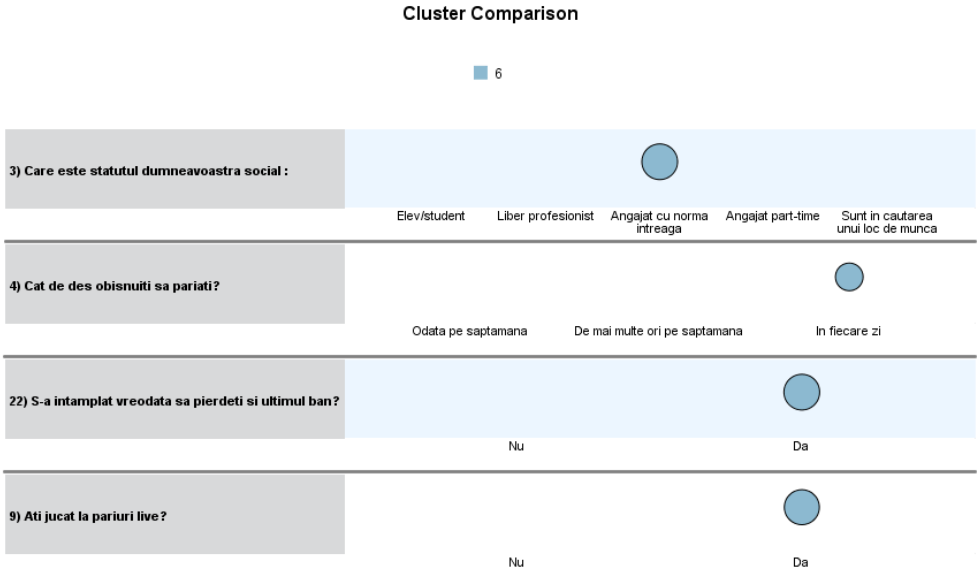


Cluster	4	6	5	2	3	1
Label						
Description	persoane angajate cu normă întreagă, pariază zilnic, nu au pierdut niciodată toți banii și au jucat la pariuri live	persoane angajate cu normă întreagă, pariază zilnic, au fost cazuri în care au pierdut toți banii și au jucat la pariuri live	elevi/studenți care pariază zilnic, nu au fost cazuri în care să piardă toți banii și au jucat la pariuri live	liber profesioniști, pariază zilnic, nu și-au pierdut toți banii la jocuri de noroc și au jucat la pariuri live	elevi/studenți, pariază odată pe săptămână, nu au pierdut toți banii, nu au jucat la pariuri live	angajați cu normă întreagă, pariază de mai multe ori pe săptămână, nu au pierdut toți banii și au jucat la pariuri live
Size	23,1% (31)	18,7% (25)	17,2% (23)	14,9% (20)	13,4% (18)	12,7% (17)
Inputs	3) Care este statutul dumneavoastra social :	3) Care este statutul dumneavoastra social :	3) Care este statutul dumneavoastra social :	3) Care este statutul dumneavoastra social :	3) Care este statutul dumneavoastra social :	3) Care este statutul dumneavoastra social :
	4) Cat de des obisnuiti sa pariati?	4) Cat de des obisnuiti sa pariati?	4) Cat de des obisnuiti sa pariati?	4) Cat de des obisnuiti sa pariati?	4) Cat de des obisnuiti sa pariati?	4) Cat de des obisnuiti sa pariati?
	22) S-a intamplat vreodata sa pierdeti si ultimul ban?	22) S-a intamplat vreodata sa pierdeti si ultimul ban?	22) S-a intamplat vreodata sa pierdeti si ultimul ban?	22) S-a intamplat vreodata sa pierdeti si ultimul ban?	22) S-a intamplat vreodata sa pierdeti si ultimul ban?	22) S-a intamplat vreodata sa pierdeti si ultimul ban?
	9) Ati jucat la pariuri live?	9) Ati jucat la pariuri live?	9) Ati jucat la pariuri live?	9) Ati jucat la pariuri live?	9) Ati jucat la pariuri live?	9) Ati jucat la pariuri live?

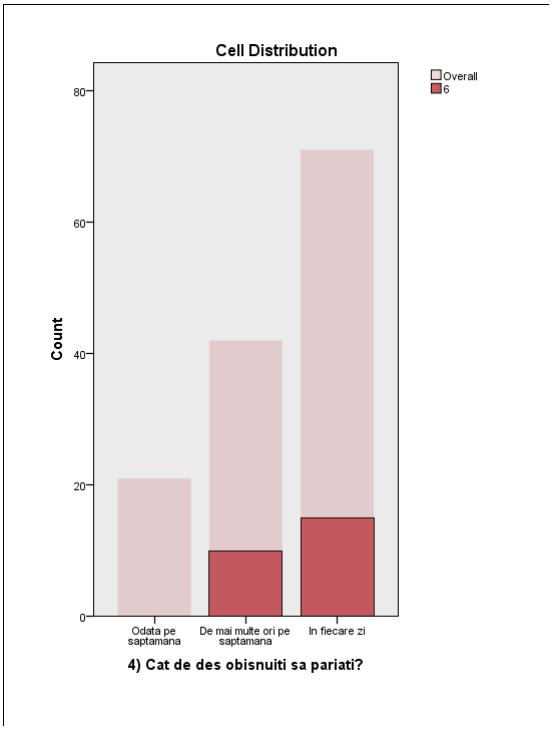
Pe rândul „Description” am notat descrierea pentru fiecare cluster reieșită din celulele din „Inpute”. Rândul „Size” ne arată dimensiunea clusterului, proporția persoanelor din grup în total. Acest tabel ordonează clusterele în ordinea descrescătoare a dimensiunii.



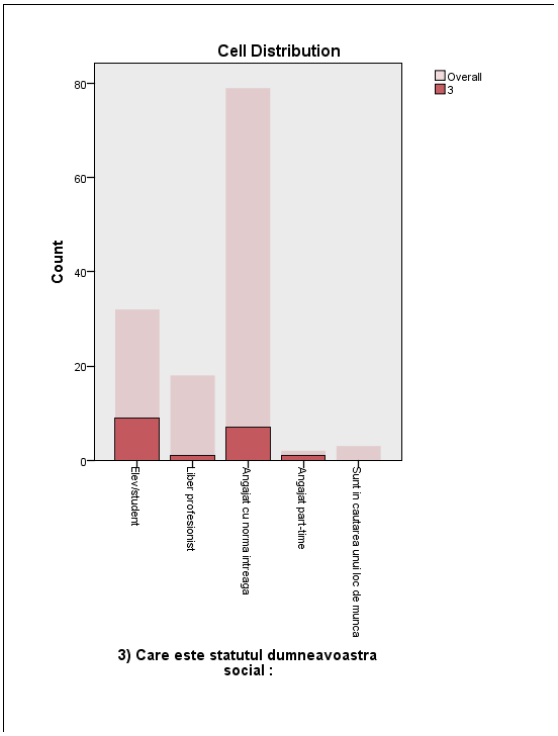
Clusterul 4 este reprezentat de persoane angajate cu normă întreagă, cu frecvență zilnică a pariurilor, acestea nu au pierdut toți banii la jocuri de noroc și au jucat la pariuri live.



În a șasea grupă avem persoane angajate cu normă întreagă, care pariază zilnic, au pierdut toți banii la jocuri de noroc și au jucat la pariuri live.



În clusterul 6, aproximativ 60% dintre respondenți pariază zilnic, iar 40% frecventează pariurile de mai multe ori pe săptămână.



În a treia grupă creată 50% dintre persoane sunt elevi sau studenți, 40% sunt angajați cu normă întreagă, 5% sunt anagajați part time și restul sunt liber profesioniști.

General Linear Model – univariate full factorial

Variabila dependentă pe care am ales-o pentru GLM este viteza maximă. Am analizat dacă aceasta urmărește legea normală de distribuție (Analyze – Frequencies).

Statistics

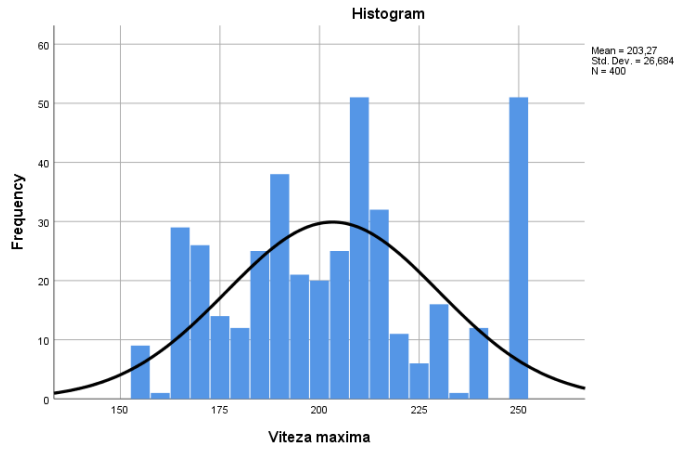
Viteza maxima

N	Valid	400
	Missing	0
Mean		203,28
Median		203,00
Std. Deviation		26,684
Skewness		,248
Std. Error of Skewness		,122
Kurtosis		-,796
Std. Error of Kurtosis		,243
Minimum		155
Maximum		250

Avem un eșantion de 400 de observații, dintre care nu avem nicio valoare izolată. Media și mediana au valori apropiate ceea ce ne indică o distribuție normală. Valorile pentru asimetrie și boltire sunt destul de aproape de zero, acesta fiind o altă caracteristică a normalității distribuției.

Pe grafic putem observa un număr mare de observații care sunt înafara curbei normale. Aceste observații nu se pot izola pentru că reprezintă 12.8% din numărul total de observații.

Variabila noastră tinde către o distribuție normală, deci putem continua analiza.



Factorii aleși pentru a continua GLM sunt:

- Preț pe intervale – variabilă ordinală cu 4 stări
- Tip motor – variabilă nominală cu 4 stări

Setările efectuate pentru GLM:

- Plots: pret_int*Motor
- EM Means – Display means for: toate opțiunile – Compare Main effects LSD
- Options: Descriptive statistics, parameter estimates, homogeneity tests, residual plot

Between-Subjects Factors

		Value Label	N
pret pe intervale	1,00	pana la 5000	66
	2,00	5000-10000	141
	3,00	10000-20000	79
	4,00	peste 20000	114
Tip motor	1	Benzina(motor aspirat)	55
	2	Benzina(motor cu turbina)	24
	3	Diesel(motor aspirat)	8
	4	Diesel(motorc cu turbina	313

Acest tabel ne arată numărul observațiilor pentru fiecare stare în parte a variabilelor factor. De exemplu, automobilele cu un preț maxim de 5000 de euro sunt în număr de 66. Iar mașinile care au motor diesel cu turbină sunt 313.

Descriptive Statistics

Dependent Variable: Viteza maxima

pret pe intervale	Tip motor	Mean	Std. Deviation	N
pana la 5000	Benzina(motor aspirat)	171,29	11,145	31
	Diesel(motor aspirat)	160,71	5,345	7
	Diesel(motorc cu turbina	184,21	15,488	28
	Total	175,65	14,976	66
5000-10000	Benzina(motor aspirat)	174,54	10,697	24
	Benzina(motor cu turbina)	195,38	18,275	8
	Diesel(motor aspirat)	174,00	.	1
	Diesel(motorc cu turbina	193,03	15,186	108
	Total	189,88	16,225	141
10000-20000	Benzina(motor cu turbina)	213,25	26,086	8

Tabelul alăturat prezintă statisticile descriptive pentru variabila dependentă, în cazul nostru viteza maximă, în funcție de variabilele factor. De exemplu, pentru mașinile cu un preț maxim de 5000 de euro cu motor aspirat pe benzină, viteza medie maximă este de 171,29 km/h. Observațiile încadrate în această categorie de abat de la medie, în medie, cu 11.145 unități.

	Diesel(motorc cu turbina	209,00	15,772	71
	Total	209,43	16,911	79
peste 20000	Benzina(motor cu turbina)	240,00	18,891	8
	Diesel(motorc cu turbina	230,93	18,853	106
	Total	231,57	18,916	114
Total	Benzina(motor aspirat)	172,71	10,972	55
	Benzina(motor cu turbina)	216,21	27,717	24
	Diesel(motor aspirat)	162,38	6,823	8
	Diesel(motorc cu turbina	208,70	24,127	313
	Total	203,27	26,684	400

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
Viteza maxima	Based on Mean	7,188	9	389	,000
	Based on Median	4,535	9	389	,000
	Based on Median and with adjusted df	4,535	9	302,728	,000
	Based on trimmed mean	7,092	9	389	,000

Formulăm ipoteza nulă: Variația de la nivelul grupelor nu se diferențiază.

Sig= 0.000 < 0.05 deci ipoteza nulă se respinge, există cel puțin 2 grupe în care variația diferă semnificativ.

Tests of Between-Subjects Effects

Dependent Variable: Viteza maxima

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	182082,432[a]	10	18208,243	69,428	,000
Intercept	2210600,628	1	2210600,628	8429,028	,000
pret_int	37516,872	3	12505,624	47,684	,000
Motor	6949,659	3	2316,553	8,833	,000
pret_int * Motor	459,263	4	114,816	,438	,781
Error	102019,318	389	262,260		
Total	16812392,000	400			
Corrected Total	284101,750	399			
a. R Squared = ,641 (Adjusted R Squared = ,632)					

Formulăm ipoteza nulă pentru fiecare variabilă în parte:

H_0: Prețul automobilelor (pe intervale) nu este semnificativ pentru model.

Sig=0.000 deci ipoteza nulă se respinge, prețul este semnificativ pentru modelul nostru.

H_0: Interacțiunea dintre preț și tipul de motor nu este semnificativ pentru model.

Sig=0,781 deci ipoteza nulă se acceptă oricare ar fi pragul de semnificativitate, deci această interacțiune nu produce efecte semnificative pentru model.

Acest tabel ne prezintă și coeficientul de determinație R^2=0,641. Deci variația vitezei maxime este explicată în proporție de 64% de variația variabilelor factor, preț și tip de motor.

Parameter Estimates

Dependent Variable: Viteza maxima

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	230,934	1,573	146,816	,000	227,841	234,026
[pret_int=1,00]	-46,720	3,441	-13,577	,000	-53,485	-39,954
[pret_int=2,00]	-37,906	2,214	-17,120	,000	-42,259	-33,553
[pret_int=3,00]	-21,934	2,484	-8,832	,000	-26,817	-17,051
[pret_int=4,00]	0[a]	.	.	.	.	.
[Motor=1]	-18,486	3,655	-5,058	,000	-25,671	-11,301
[Motor=2]	9,066	5,938	1,527	,128	-2,608	20,740
[Motor=3]	-19,028	16,269	-1,170	,243	-51,014	12,959
[Motor=4]	0[a]	.	.	.	.	.
[pret_int=1,00] * [Motor=1]	5,562	5,584	,996	,320	-5,417	16,541
[pret_int=1,00] * [Motor=3]	-4,472	17,650	-,253	,800	-39,173	30,229
[pret_int=1,00] * [Motor=4]	0[a]	.	.	.	.	.
[pret_int=2,00] * [Motor=1]	0[a]	.	.	.	.	.
[pret_int=2,00] * [Motor=2]	-6,719	8,394	-,800	,424	-23,223	9,785
[pret_int=2,00] * [Motor=3]	0[a]	.	.	.	.	.
[pret_int=2,00] * [Motor=4]	0[a]	.	.	.	.	.
[pret_int=3,00] * [Motor=2]	-4,816	8,470	-,569	,570	-21,468	11,836
[pret_int=3,00] * [Motor=4]	0[a]	.	.	.	.	.
[pret_int=4,00] * [Motor=2]	0[a]	.	.	.	.	.
[pret_int=4,00] * [Motor=4]	0[a]	.	.	.	.	.

Tabelul de mai sus estimează parametri modelului și permite efectuarea de comparații între viteza maximă în funcție de variabilele factor. De exemplu, mașinile cu un preț maxim de 5000 de euro au o

viteză maximă, în medie, mai mică față de automobilele cu un preț de peste 20.000 de euro, aceasta fiind categoria de referință. Sig= 0,000 deci avem o diferență semnificativă între cele 2 grupe.

Pairwise Comparisons

Dependent Variable: Viteza maxima

(I) pret pe intervale	(J) pret pe intervale	Mean Difference (I-J)	Std. Error	Sig.[d]	95% Confidence Interval for Difference[d]	
					Lower Bound	Upper Bound
pana la 5000	5000-10000	-12,163[*,b]	5,042	,016	-22,075	-2,251
	10000-20000	-39,052[*,b,c]	3,907	,000	-46,733	-31,371
	peste 20000	-63,394[*,b,c]	3,868	,000	-70,998	-55,790
5000-10000	pana la 5000	12,163[*,c]	5,042	,016	2,251	22,075
	10000-20000	-26,889[*,c]	5,329	,000	-37,365	-16,412
	peste 20000	-51,231[*,c]	5,300	,000	-61,651	-40,811
10000-20000	pana la 5000	39,052[*,b,c]	3,907	,000	31,371	46,733
	5000-10000	26,889[*,b]	5,329	,000	16,412	37,365
	peste 20000	-24,342[*,b,c]	4,235	,000	-32,668	-16,016
peste 20000	pana la 5000	63,394[*,b,c]	3,868	,000	55,790	70,998
	5000-10000	51,231[*,b]	5,300	,000	40,811	61,651
	10000-20000	24,342[*,b,c]	4,235	,000	16,016	32,668

Acest tabel ne prezintă comparația între automobilele cu prețuri diferite în funcție de viteza maximă.

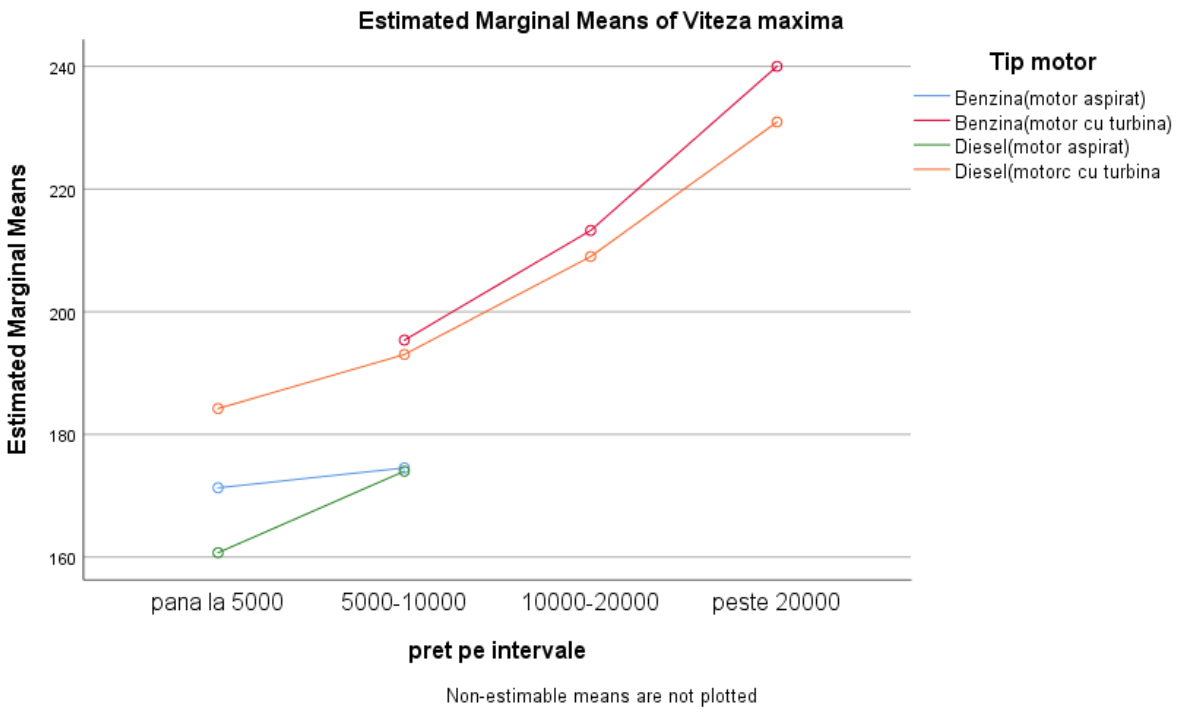
Se formulează ipotezele nule pentru fiecare categorie: Nu există o diferență semnificativă între viteza maximă a mașinilor cu preț maxim de 5000 de euro față de cele cu preț cuprins între 5000-1000 de euro. Sig=0.016 < 0.05 deci ipoteza nulă se respinge, deci diferența este semnificativă.

Pairwise Comparisons

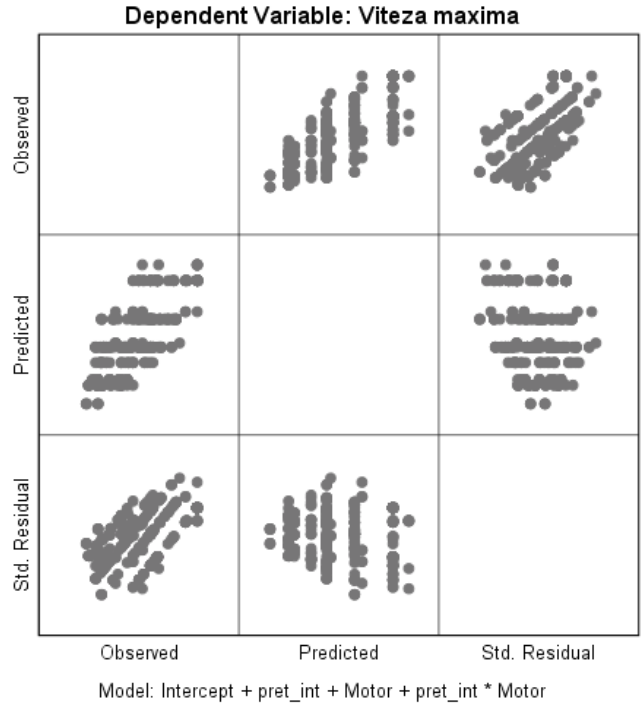
Dependent Variable: Viteza maxima

(I) Tip motor	(J) Tip motor	Mean Difference (I-J)	Std. Error	Sig.[d]	95% Confidence Interval for Difference[d]	
					Lower Bound	Upper Bound
Benzina(motor aspirat)	Benzina(motor cu turbina)	-43,292[*,b,c]	3,972	,000	-51,101	-35,484
	Diesel(motor aspirat)	5,559[b,c]	8,932	,534	-12,002	23,120
	Diesel(motorc cu turbina	-31,378[*,b]	2,443	,000	-36,182	-26,574
Benzina(motor cu turbina)	Benzina(motor aspirat)	43,292[*,b,c]	3,972	,000	35,484	51,101
	Diesel(motor aspirat)	48,851[*,b,c]	9,266	,000	30,633	67,069
	Diesel(motorc cu turbina	11,914[*,b]	3,471	,001	5,089	18,739
Diesel(motor aspirat)	Benzina(motor aspirat)	-5,559[b,c]	8,932	,534	-23,120	12,002
	Benzina(motor cu turbina)	-48,851[*,b,c]	9,266	,000	-67,069	-30,633
	Diesel(motorc cu turbina	-36,937[*,b]	8,721	,000	-54,083	-19,791
Diesel(motorc cu turbina	Benzina(motor aspirat)	31,378[*,c]	2,443	,000	26,574	36,182
	Benzina(motor cu turbina)	-11,914[*,c]	3,471	,001	-18,739	-5,089
	Diesel(motor aspirat)	36,937[*,c]	8,721	,000	19,791	54,083

Aici putem compara viteza maximă a automobilelor în funcție de tipul motorului. Putem observa că diferența dintre viteza maximă a mașinilor cu motor diesel aspirat față de cele cu motor pe benzină aspirat este nesemnificativă.



Graficul ne arată dacă avem interacțiuni între grupele create.



Din acest grafic putem spune că valorile observate și cele reziduale au o legătură de intensitate medie, deci nu am reușit să explicăm în totalitate variația vitezei maxime prin variația factorilor aleși.

General Linear Model – univariate custom factor

Pentru această analiză am continuat cu viteza maximă ca variabilă dependentă. Ca variabile factor am ales:

- Tip motor – variabilă nominală cu 4 stări
- Preț pe intervale – variabilă ordinală cu 4 stări
- Cost de întreținere – variabilă ordinală cu 3 stări
- Tracțiunea mașinii – variabilă nominală cu 3 stări

Inițial am rulat analiza folosind aceste variabile și am avut un coeficient de determinație R^2=0,691, adică variația vitezei maxime depinde în proporție de 69% de variația factorilor aleși.

Pentru a îmbunătăți calitatea modelului am adăugat ca variabilă covariată, puterea. În urma rulării analizei folosind această nouă variabilă avem o calitate mai bună a modelului cu un R^2=0,873.

Datorită numărului mare de combinații între observații în funcție de variabilele factor, voi atașa doar o parte din statisticile descriptive.

Descriptive Statistics

Dependent Variable: Viteza maxima

Tip motor	pret pe intervale	Cost intretinere	Tractiunea masinii	Mean	Std. Deviation	N
Diesel(motorc cu turbina	pana la 5000	Scazut	Fata	172,25	10,500	4
			Total	172,25	10,500	4
		Mediu	Fata	184,62	15,513	21
			Total	184,62	15,513	21
		Ridicat	Fata	197,33	10,970	3
			Total	197,33	10,970	3
		Total	Fata	184,21	15,488	28
			Total	184,21	15,488	28
	5000-10000	Scazut	Fata	186,45	9,903	11
			4x4	168,00	.	1
			Total	184,92	10,841	12
		Mediu	Fata	190,46	15,009	70
			Spate	203,75	5,175	8
			4x4	202,67	12,702	3
			Total	192,22	14,858	81
		Ridicat	Fata	207,27	15,087	11
			4x4	194,50	11,000	4
			Total	203,87	14,923	15
		Total	Fata	191,99	15,503	92
			Spate	203,75	5,175	8
			4x4	194,25	15,059	8
			Total	193,03	15,186	108

Deci, luând în considerare aceste statistici, rezultă că viteza maximă a unei mașini cu motor diesel cu turbină cu cost de întreținere mediu cu un preț cuprins între 5000-10.000 de euro cu tracțiune față

este, în medie, 190,46 km/h. În timp ce un automobil cu același tip de motor, cu un preț cuprins în același interval și aceeași tracțiune, dar cu un cost ridicat de întreținere are o viteză maximă medie de 207 km/h.

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Viteza maxima

F	df1	df2	Sig.
13,341	39	360	,000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Motor + pret_int + Intretinere + Tractiune + Putere

Cu ajutorul testului Levene testăm ipoteza nulă că variația de la nivelul grupelor nu se diferențiază. Sig=0,000 mai mic decât orice prag de risc, deci ipoteza nulă se respinge, adică există cel puțin 2 grupe în care variația diferă semnificativ.

Tests of Between-Subjects Effects

Dependent Variable: Viteza maxima

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	248107,828[a]	11	22555,257	243,137	,000
Intercept	187799,992	1	187799,992	2024,408	,000
Motor	964,262	3	321,421	3,465	,016
pret_int	717,177	3	239,059	2,577	,053
Intretinere	3646,801	2	1823,401	19,656	,000
Tractiune	9908,892	2	4954,446	53,407	,000
Putere	51749,536	1	51749,536	557,839	,000
Error	35993,922	388	92,768		
Total	16812392,000	400			
Corrected Total	284101,750	399			

Pentru fiecare variabilă formulăm ipoteza nulă:

De exemplu, H_0: Tipul motorului nu este semnificativ pentru model.

Sig=0,016 < 0,05 deci ipoteza nulă se respinge, adică tipul motorului este semnificativ în a explica viteza maximă a automobilelor.

Parameter Estimates

Dependent Variable: Viteza maxima

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	125,232	4,639	26,995	,000	116,111	134,353
[Motor=1]	-5,018	1,717	-2,922	,004	-8,394	-1,642
[Motor=2]	-,045	2,109	-,021	,983	-4,192	4,101
[Motor=3]	-7,768	3,742	-2,076	,039	-15,124	-,411
[Motor=4]	0[a]	.	.	.	.	.
[pret_int=1,00]	-5,677	2,647	-2,145	,033	-10,880	-,474
[pret_int=2,00]	-4,190	2,135	-1,962	,050	-8,388	,008
[pret_int=3,00]	-,258	1,733	-,149	,882	-3,664	3,149
[pret_int=4,00]	0[a]	.	.	.	.	.
[Intretinere=1]	3,747	2,318	1,616	,107	-,811	8,305
[Intretinere=2]	9,159	1,629	5,621	,000	5,955	12,362
[Intretinere=3]	0[a]	.	.	.	.	.
[Tractiune=1]	13,649	1,647	8,288	,000	10,411	16,887
[Tractiune=2]	17,281	1,971	8,767	,000	13,406	21,157
[Tractiune=3]	0[a]	.	.	.	.	.
Putere	,431	,018	23,619	,000	,395	,467

a. This parameter is set to zero because it is redundant.

Tabelul de mai sus estimează parametri modelului și permite efectuarea de comparații între viteza maximă în funcție de variabilele factor. De exemplu, mașinile cu motor aspirat pe benzină au o

viteză maximă medie mai mică decât cele cu motor diesel cu turbină, având o diferență semnificativă (sig=0,000 < 0,05). mașinile cu motor cu turbină pe benzină au o viteză maximă medie mai mică decât cele cu motor cu turbină pe diesel, diferența nu este semnificativă (sig=0,983).

În continuare avem comparații între viteza maximă în funcție de fiecare variabilă factor.

- Tip motor

Estimates

Dependent Variable: Viteza maxima

Tip motor	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Benzina(motor aspirat)	198,724[a]	1,609	195,561	201,888
Benzina(motor cu turbina)	203,697[a]	2,239	199,295	208,099
Diesel(motor aspirat)	195,975[a]	3,631	188,836	203,113
Diesel(motor cu turbina)	203,742[a]	,905	201,964	205,521

Din acest tabel putem concluziona că mașinile care au motor cu turbină, indiferent de tipul de combustibil sunt cele mai rapide, în timp ce automobilele cu motor diesel aspirat au cea mai mică viteză maximă dintre toate.

Pairwise Comparisons

Dependent Variable: Viteza maxima

(I) Tip motor	(J) Tip motor	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Benzina(motor aspirat)	Benzina(motor cu turbina)	-4,973	2,692	,065	-10,265	,320
	Diesel(motor aspirat)	2,750	3,688	,456	-4,501	10,001
	Diesel(motorc cu turbina	-5,018*	1,717	,004	-8,394	-1,642
Benzina(motor cu turbina)	Benzina(motor aspirat)	4,973	2,692	,065	-,320	10,265
	Diesel(motor aspirat)	7,723	4,303	,073	-,738	16,183
	Diesel(motorc cu turbina	-,045	2,109	,983	-4,192	4,101
Diesel(motor aspirat)	Benzina(motor aspirat)	-2,750	3,688	,456	-10,001	4,501
	Benzina(motor cu turbina)	-7,723	4,303	,073	-16,183	,738
	Diesel(motorc cu turbina	-7,768*	3,742	,039	-15,124	-,411
Diesel(motorc cu turbina	Benzina(motor aspirat)	5,018*	1,717	,004	1,642	8,394
	Benzina(motor cu turbina)	,045	2,109	,983	-4,101	4,192
	Diesel(motor aspirat)	7,768*	3,742	,039	,411	15,124

Acest tabel permite comparația între viteza maximă în funcție de tipul de motor al mașinii. Formulăm ipoteza nulă: Nu există o diferență semnificativă între viteza maximă a mașinilor cu motor aspirat pe benzină și motor cu turbină pe benzină. Pentru un prag de risc de 5%, ipoteza

s-ar accepta la limită, dar pentru un prag de 10% aceasta se respinge, adică diferența este semnificativă.

- Pret pe intervale

Estimates

Dependent Variable: Viteza maxima

pret pe intervale	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
pana la 5000	197,389[a]	1,908	193,638	201,139
5000-10000	198,876[a]	1,691	195,552	202,199
10000-20000	202,808[a]	1,683	199,500	206,117
peste 20000	203,066[a]	1,780	199,565	206,566

De aici putem concluziona că pe măsură ce crește prețul, crește și viteza maximă medie a mașinii.

Pairwise Comparisons

Dependent Variable: Viteza maxima

(I) pret pe intervale	(J) pret pe intervale	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
pana la 5000	5000-10000	-1,487	1,612	,357	-4,657	1,683
	10000-20000	-5,419*	2,145	,012	-9,637	-1,201
	peste 20000	-5,677*	2,647	,033	-10,880	-,474
5000-10000	pana la 5000	1,487	1,612	,357	-1,683	4,657
	10000-20000	-3,932*	1,607	,015	-7,091	-,774
	peste 20000	-4,190	2,135	,050	-8,388	,008
10000-20000	pana la 5000	5,419*	2,145	,012	1,201	9,637
	5000-10000	3,932*	1,607	,015	,774	7,091
	peste 20000	-,258	1,733	,882	-3,664	3,149
peste 20000	pana la 5000	5,677*	2,647	,033	,474	10,880
	5000-10000	4,190	2,135	,050	-,008	8,388
	10000-20000	,258	1,733	,882	-3,149	3,664

Cu un prag de risc de 5% putem spune că nu există o diferență semnificativă între viteza maximă a mașinilor cu un preț cuprins în intervalul 5000-10.000 de euro față de cea a mașinilor cu un preț sub 5000 de euro.

- Cost de întreținere

Estimates

Dependent Variable: Viteza maxima

Cost intretinere	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Scazut	199,979[a]	1,937	196,171	203,788
Mediu	205,391[a]	1,410	202,619	208,164
Ridicat	196,233[a]	1,661	192,967	199,498

Mașinile cu un cost mediu de întreținere au, în medie, cea mai mare viteză maximă.

Pairwise Comparisons

Dependent Variable: Viteza maxima

(I) Cost intretinere	(J) Cost intretinere	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Scazut	Mediu	-5,412*	1,746	,002	-8,845	-1,979
	Ridicat	3,747	2,318	,107	-,811	8,305
Mediu	Scazut	5,412*	1,746	,002	1,979	8,845
	Ridicat	9,159*	1,629	,000	5,955	12,362
Ridicat	Scazut	-3,747	2,318	,107	-8,305	,811
	Mediu	-9,159*	1,629	,000	-12,362	-5,955

Având în vedere costul de întreținere al automobilelor, observăm că nu avem o diferență semnificativă între viteza maximă a celor cu un cost scăzut față de cele cu un cost ridicat.

- Tracțiunea mașinii

Estimates

Dependent Variable: Viteza maxima

Tractiunea masinii	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Fata	203,874[a]	1,269	201,378	206,369
Spate	207,506[a]	2,103	203,371	211,641

4x4	190,224[a]	1,544	187,188	193,261

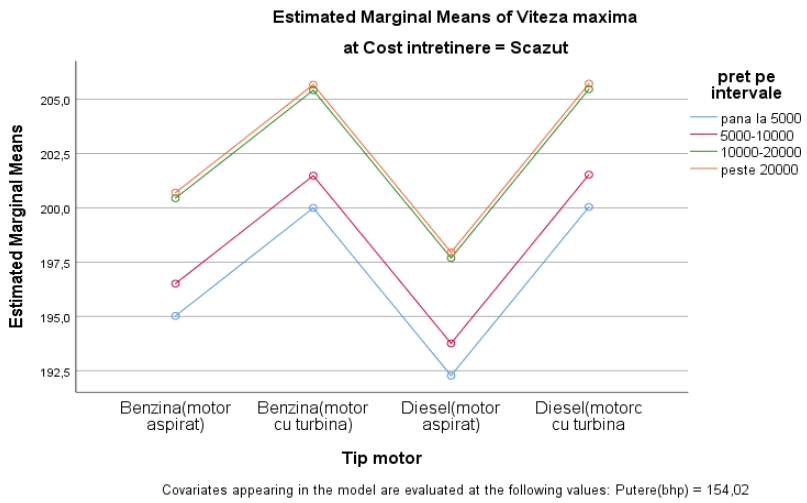
Mașinile cu tracțiune spate au, în medie, cea mai mare viteză maximă față de cele 4x4 la care viteza maximă este mai mică cu aproape 30 km/h.

Pairwise Comparisons

Dependent Variable: Viteza maxima

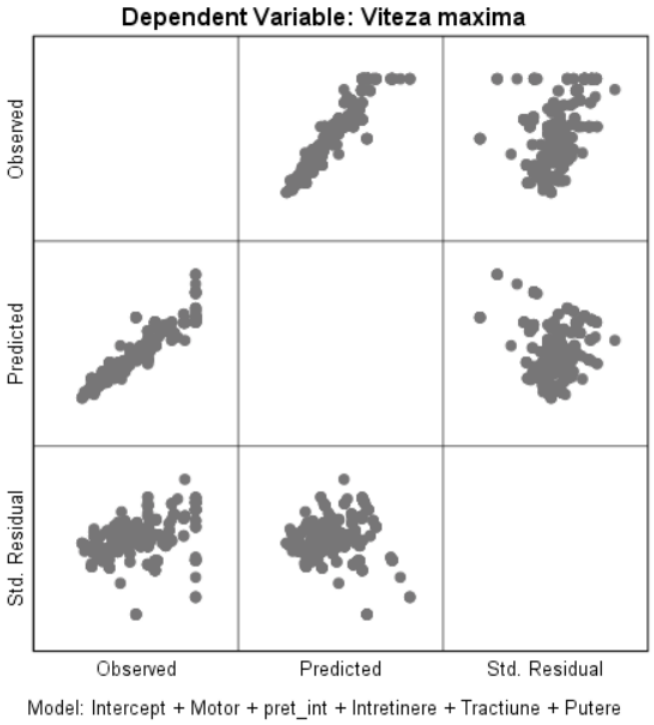
(I) Tractiunea masinii	(J) Tractiunea masinii	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Fata	Spate	-3,632	2,058	,078	-7,678	,414
	4x4	13,649*	1,647	,000	10,411	16,887
Spate	Fata	3,632	2,058	,078	-,414	7,678
	4x4	17,281*	1,971	,000	13,406	21,157
4x4	Fata	-13,649*	1,647	,000	-16,887	-10,411
	Spate	-17,281*	1,971	,000	-21,157	-13,406

Tabelul ne arată că avem diferențe semnificative între elementele perechilor, exceptând viteza maximă medie a mașinilor cu tracțiune spate care nu se diferențiază de cele cu tracțiune față.



Din acest grafic putem spune că dintre mașinile cu cost scăzut de întreținere, cea mai mare viteză maximă o au cele cu motor cu turbină, indiferent de combustibil și de intervalele de preț.

Iar cea mai mică viteză maximă o au automobilele cu motor aspirat pe diesel cu un preț maxim de 5000 de euro



În concluzie, punctele din graficul între valorile observate și cele previzionate sunt grupate, prin urmare avem un model bun. Iar cele dintre reziduuri și valorile observate sunt destul de împrăștiate, ceea ce înseamnă că în reziduuri nu mai eixstă informație care explică viteza maximă a automobilelor.

Analiza regresională

Pentru această analiză am ales să folosesc ca variabilă dependentă prețul de vânzare al unor apartamente. Am testat dacă variabila are o distribuție normală:

Statistics

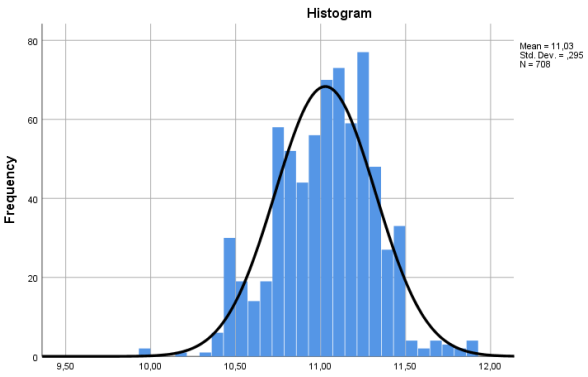
Selling price

N	Valid	708
	Missing	0
Mean		64236,0410
Median		62250,0000
Std. Deviation		18971,58358
Skewness		,874
Std. Error of Skewness		,092
Kurtosis		2,081
Std. Error of Kurtosis		,183

Media și mediana au valori apropiate, ceea ce indică o distribuție normală. Însă abaterea standard, asimetria și boltirea au valori prea mari pentru ca variabila să urmeze o lege normală de probabilitate. Pe histogramă se vede clar acest lucru.

În acest caz am ales să logaritmez variabila și să rulez iar analiza.

N	Valid	708
	Missing	0
Mean		11,0275
Median		11,0389
Std. Deviation		,29523
Skewness		-,189
Std. Error of Skewness		,092
Kurtosis		,236
Std. Error of Kurtosis		,183



Acum putem spune că variabila tinde către o distribuție normală.

Variabilele independente alese au fost:

- Vârsta imobilului
- Distanța față de centru
- Suprafața utilă

Descriptive Statistics

	Mean	Std. Deviation	N
lnPretVanzare	11,0275	,29523	708
Age	25,4689	13,48215	708
Distance from downtown	2,8894	1,80311	708
Useful area	40,7528	13,15597	708

În acest tabel ne sunt prezentate media, abaterea standard și numărul de observații pentru fiecare variabilă.

Correlations

		lnPretVanzare	Age	Distance from downtown	Useful area
Pearson Correlation	lnPretVanzare	1,000	-,281	-,062	,857
	Age	-,281	1,000	-,055	-,292
	Distance from downtown	-,062	-,055	1,000	-,142
	Useful area	,857	-,292	-,142	1,000
Sig. (1-tailed)	lnPretVanzare	.	,000	,050	,000
	Age	,000	.	,073	,000
	Distance from downtown	,050	,073	.	,000
	Useful area	,000	,000	,000	.
N	lnPretVanzare	708	708	708	708
	Age	708	708	708	708
	Distance from downtown	708	708	708	708
	Useful area	708	708	708	708

Acest tabel ne arată corelația dintre variabilele alese pentru analiză. Prețul de vânzare și vârsta imobilului prezintă o legătură inversă, de intensitate slabă. În aceeași situație se află și distanța față de centru, iar legătura dintre prețul de vânzare și suprafața utilă este una directă și de intensitate mare.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	10,232	,028		361,294	,000					
	Age	-,001	,000	-,028	-1,374	,170	-,281	-,052	-,026	,905	1,105
	Distance from downtown	,009	,003	,058	2,960	,003	-,062	,111	,057	,970	1,031
	Useful area	,019	,000	,857	41,913	,000	,857	,845	,808	,890	1,124

a. Dependent Variable: lnPretVanzare

Statistica VIF ne arată dacă avem multicoliniaritate între variabilele independente. Toate valorile sunt mai mici de 3, deci nu avem multicoliniaritate între factori (sau chiar dacă există, nu produce efecte semnificative).

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,859[a]	,738	,737	,15137	,738	661,851	3	704	,000	1,764

a. Predictors: (Constant), Useful area, Distance from downtown, Age

b. Dependent Variable: lnPretVanzare

Atât valorea lui R^2, cât și cea a lui R^2 indică faptul că avem un model valid pentru a explica variația prețului apartamentelor.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	45,494	3	15,165	661,851	,000[b]
	Residual	16,130	704	,023		
	Total	61,624	707			

a. Dependent Variable: lnPretVanzare

b. Predictors: (Constant), Useful area, Distance from downtown, Age

Formulăm ipoteza nulă: $R^2=0$ sau implicit F=0, adică modelul nostru nu este reprezentativ.

Se respinge ipoteza nulă indiferent de pragul de risc, deci modelul nostru explică bine variația prețului de vânzare prin acești predictori.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	10,232	,028		361,294	,000					
	Age	-,001	,000	-,028	-1,374	,017	-,281	-,052	-,026	,905	1,105
	Distance from downtown	,009	,003	,058	2,960	,003	-,062	,111	,057	,970	1,031
	Useful area	,019	,000	,857	41,913	,000	,857	,845	,808	,890	1,124

a. Dependent Variable: lnPretVanzare

Pentru fiecare coeficient și constantă se formulează ipoteza nulă: acești coeficienți sunt egali cu zero, adică nu sunt semnificativi pentru model.

Se aplică testul t pentru semnificația unei medii. Pentru un prag de risc de 5% toți coeficienții și constanta sunt semnificativi pentru model, adică nu sunt egali cu zero.

Sensul legăturii dintre variabila dependentă și cele independente este dată de semnul coeficientului Beta nestandardizat. Între prețul de vânzare al apartamentelor și vechime avem o legătură inversă,

adică crește vechimiea imobilului, va scădea prețul. Totuși, avem o legătură directă între suprafața utilă și preț, dacă crește suprafața apartamentului, va crește și prețul.

Dacă ordonăm coeficienții Beta standardizați în modul vom avea importanța predictorilor măsurată prin efect.

I. Suprafața utilă – cel mai important predictor
II. Distanța față de centru
III. Vechimea

Ecuația de regresie este: Ymediu = 10,232 – 0,001*Vechime + 0,009*Distanta fata de centru + 0,019*Suprafata utila

Residuals Statistics[a]

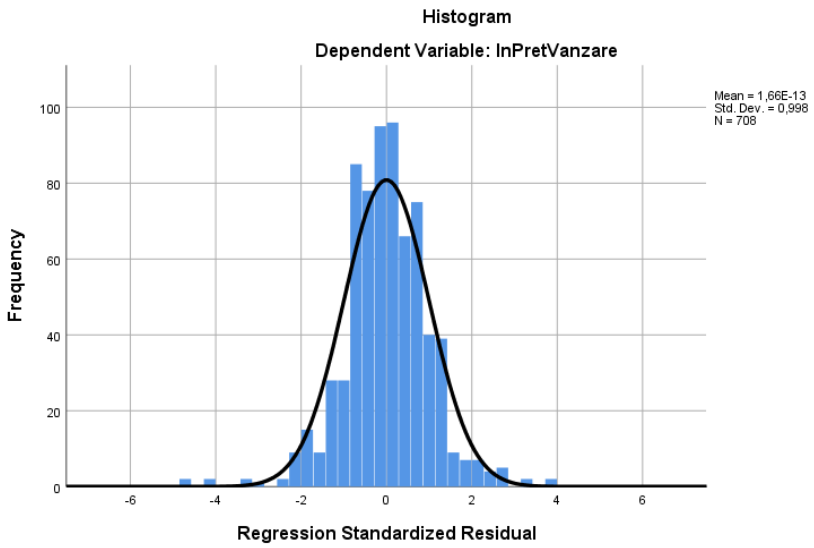
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	10,4469	12,6488	11,0275	,25367	708
Residual	-,73036	,57016	,00000	,15105	708
Std. Predicted Value	-2,289	6,391	,000	1,000	708
Std. Residual	-4,825	3,767	,000	,998	708

a. Dependent Variable: lnPretVanzare

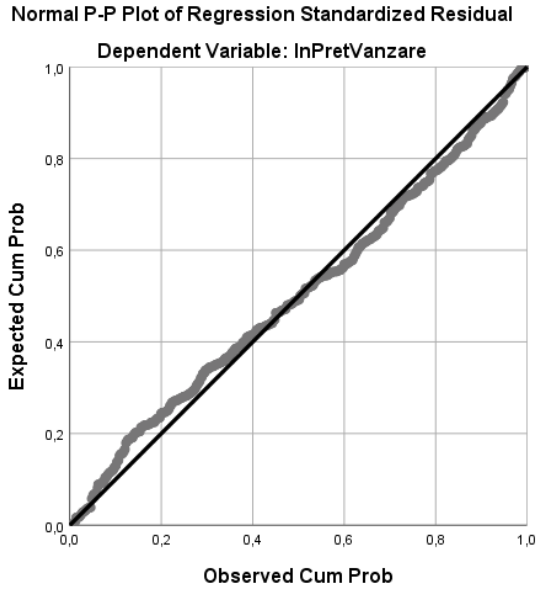
Putem spune că media valorilor previzionate în urma modelului este egală cu media valorilor observate.

Media reziduurilor (valoare reală – valoarea prezisă) este zero.

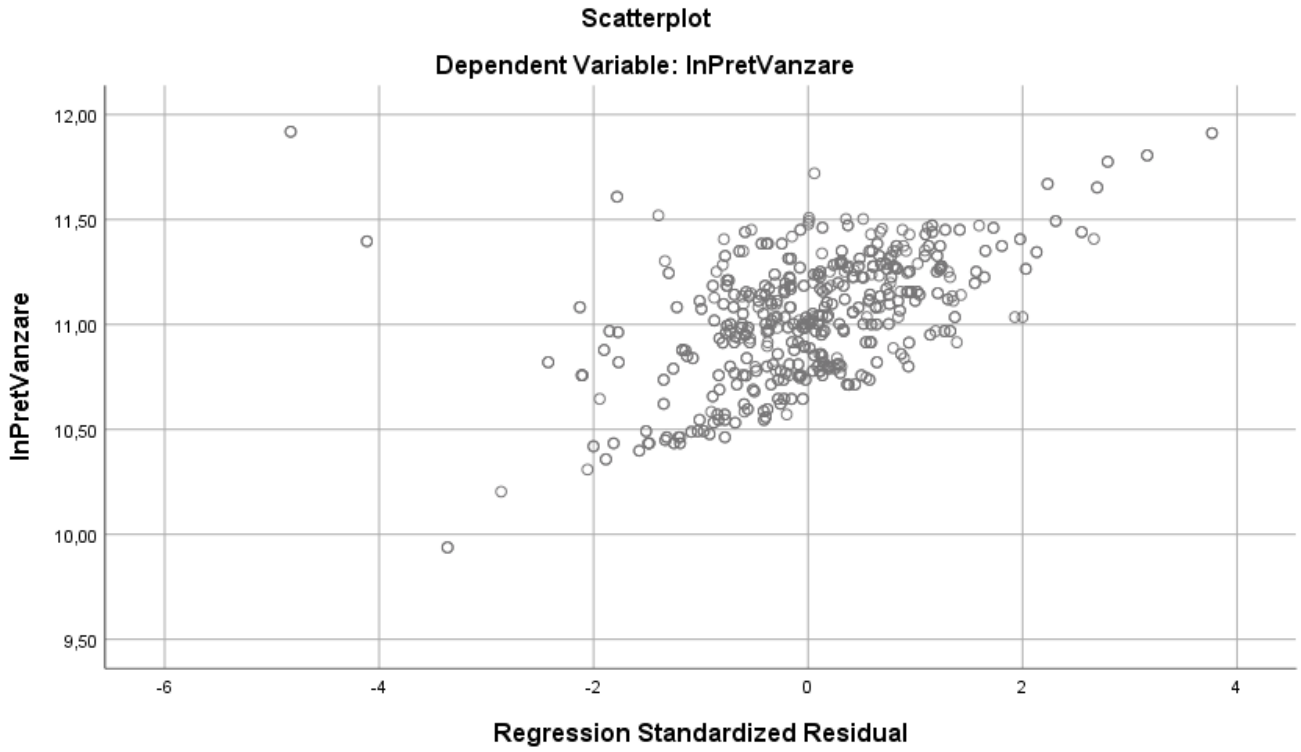
Pe al treilea rând avem statisticile pentru valorile previzionate standardizate cu medie zero și abatere 1.



Un model este reprezentativ și explică cât mai mult din variația variabilei dependente dacă distribuția reziduurilor tinde spre una normală



Faptul că avem o distribuție normală în reziduuri se vede și în acest grafic deoarece linia curbă se pliază pe trend.



Putem spune că nu avem corelație în reziduuri pentru că punctele de pe grafic sunt relativ împrăștiate, deci avem un model bun.