# TwoStep Cluster

Dinu Ioana
Icușcă Ana-Maria
Suciu Andreea

# Caracteristici

**TWO STEP CLUSTER**

- Este foarte eficientă în clasificarea bazelor mari de date
- Poate utiliza în analiză atât variabile categoriale (ordinale, nominale), cât și continue (numerice- scale)
- Prezintă avantaje în comparație cu metodele tradiționale de clustrizare (K-means și Hierarchical Cluster)

TWO STEP CLUSTER

# Analiza pieței de autovehicule

- **Analiza:** Evaluarea pieței de autovehicule (152 de mașini și camioane) în funcție de caracteristicile fizice, preț și model prin metoda *two step cluster.*

- **Scop:**

Analiza a  fost derulată pentru ca producătorii de mașini să cunoască competiția existentă pentru autovehiculele pe care urmează să le fabrice;

Autovehiculele au fost grupate în 3 clustere pentru a ilustra ce preț și ce caracteristici definesc un anumit tip de autovehicul;

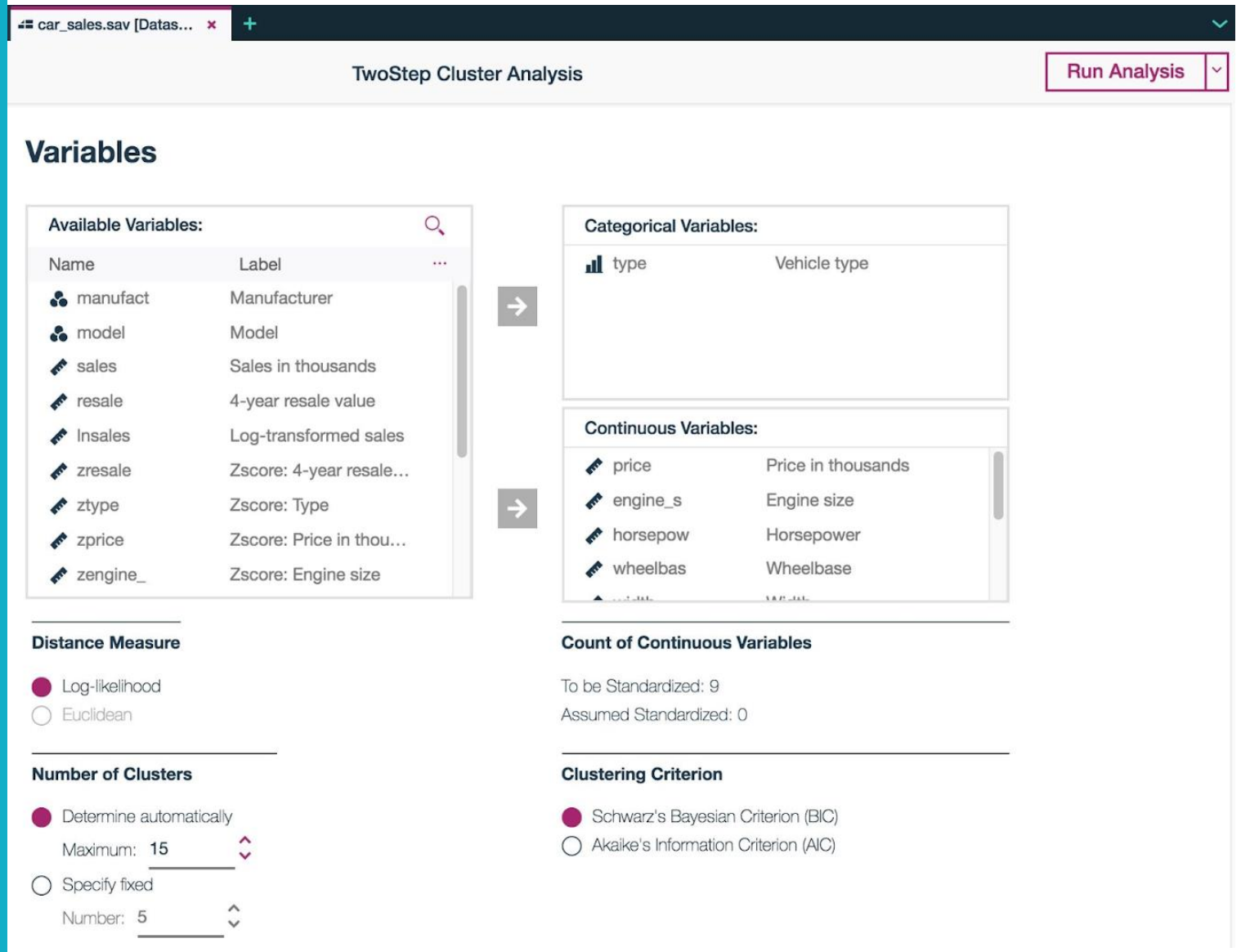
- În analiză s-au folosit:

- o variabilă ordinală (categorială);
- 9 variabile scale (continue).

- Măsurarea distanței: log-likelihood

- Criteriul de Clusterizare este cel **Bayesian.**

- Se utilizează procedura TwoStep Cluster pentru a grupa automobilele în funcţie de Tip, prețurile și proprietățile lor fizice.



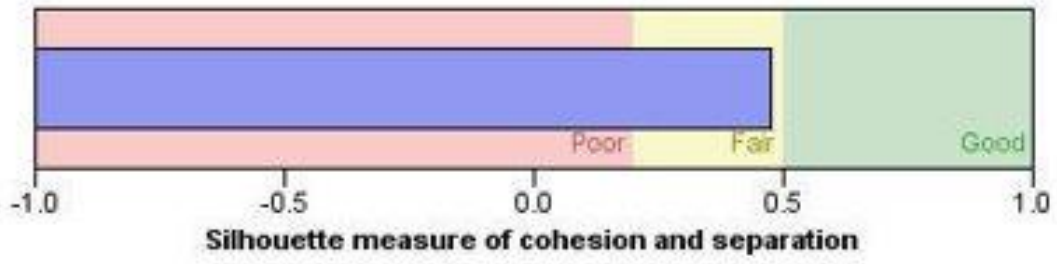
# Auto-Clustering

- Prin metoda Auto-Clustering s-a stabilit că numărul optim de clustere este 3, iar tabelul Cluster Quality confirmă eficiența lor.

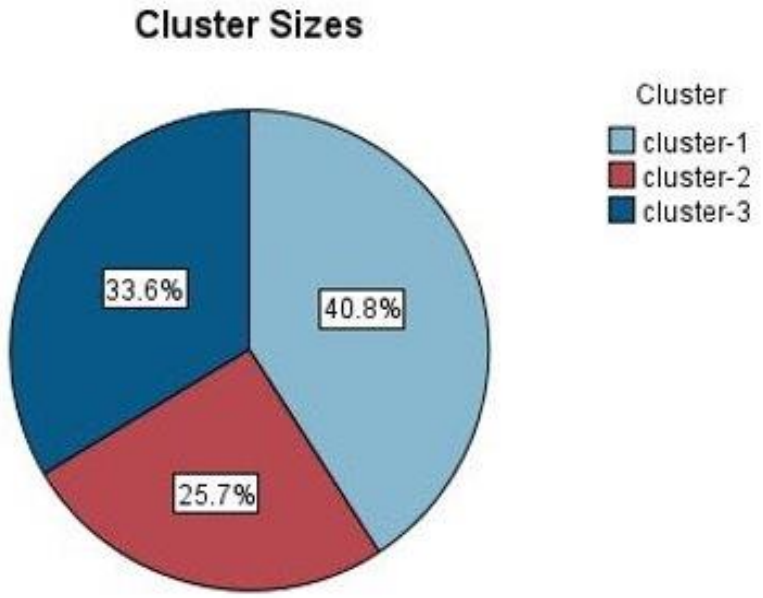
**Model Summary**

| | |
|---|---|
| **Algorithm** | TwoStep |
| **Inputs** | 10 |
| **Clusters** | 3 |

**Cluster Quality**



# Cluster Sizes




*În diagramă sunt prezentate frecvențele fiecărui cluster:*

- **40.8%** (62) din înregistrări sunt atribuite primului cluster

-**25.7%** (39) din înregistrări sunt atribuite celui de-al doilea cluster

-**33.6%** (51) din înregistrări sunt atribuite celui de-al treilea cluster

| | |
|---|---|
| **Size of Smallest Cluster** | 39 (25.7%) |
| **Size of Largest Cluster** | 62 (40.8%) |
| **Ratio of Sizes: Largest Cluster to Smallest Cluster** | 1.59 |

## Clusters

Feature Importance

1.0 0.9 0.8 0.7 0.6 0.5 0.4

| Cluster | 1 | 3 | 2 |
|---|---|---|---|
| Label | | | |
| Description | Automobile mici, ieftine, cu o capacitate mica a rezervorului si eficiente dpdv al consumului de combustibil | Automobile scumpe, mari, cu o eficiență medie în ceea ce privește capacitatea rezervorului și consumul de combustibil | Camioane la un preț mediu de dimensiuni medii, grele, cu cea mai mică eficiență a combustibilului și cel mai mare rezervor |
| Size | 40.8% (62) | 33.6% (51) | 25.7% (39) |
| Features | Vehicle type Automobile (98.4%) | Vehicle type Automobile (100.0%) | Vehicle type Truck (100.0%) |
| | Curb weight 2.84 | Curb weight 3.58 | Curb weight 3.97 |

| Features | Vehicle type Automobile (98.4%) | Vehicle type Automobile (100.0%) | Vehicle type Truck (100.0%) |
|---|---|---|---|
| | Curb weight 2.84 | Curb weight 3.58 | Curb weight 3.97 |
| | Fuel efficiency 27.24 | Fuel efficiency 23.02 | Fuel efficiency 19.51 |
| | Fuel capacity 15.00 | Fuel capacity 18.40 | Fuel capacity 22.10 |
| | Engine size 2.20 | Engine size 3.70 | Engine size 3.60 |
| | Horsepower 143.24 | Horsepower 232.96 | Horsepower 187.92 |
| | Width 68.50 | Width 72.90 | Width 72.70 |
| | Wheelbase 102.60 | Wheelbase 109.00 | Wheelbase 113.00 |
| | Length 178.20 | Length 194.70 | Length 191.10 |
| | Price in thousands 19.62 | Price in thousands 37.30 | Price in thousands 26.56 |

TWO STEP CLUSTER

# Concluzia

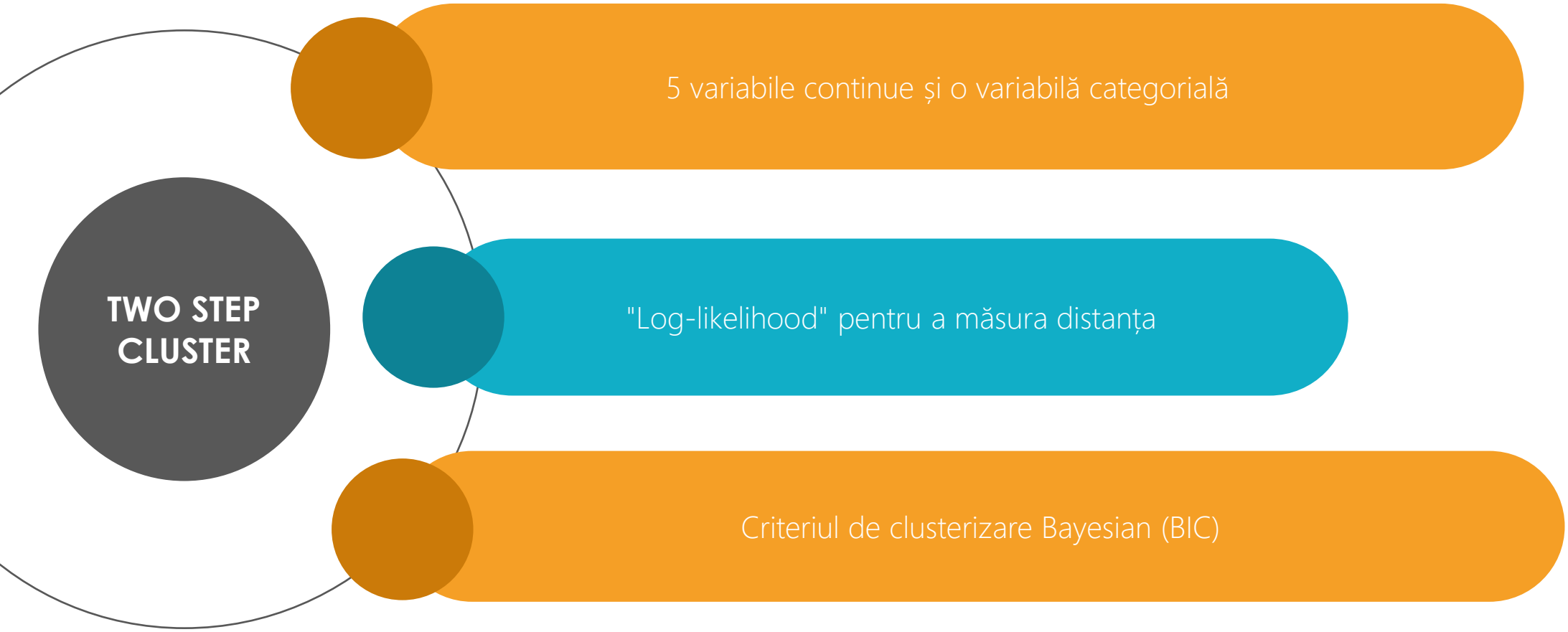
Folosind procedura TwoStep Cluster Analysis,
am separat autovehiculele în trei categorii destul de mari.

Testul ajută producătorii de autovehicule să observe concurența, analizând care sunt caracteristicile pentru un anumit tip de autovehicul (in cazul nostru automobilele mici, mari si camioanele) înainte de stabilirea unui anumit preț.

TWO STEP CLUSTER

# Segmentarea celor mai mari companii din Macedonia

Scopul acestui studiu de caz este de a analiza situația companiilor macedoniene, de a-i informa pe potențialii investitori străini, precum și de a înștiința guvernul cu privire la posibilele investiții și atrageri de fonduri pentru infrastructură, educație, industrie. Anual, Registrul Central al Republicii Macedoniene prezintă clasamentul celor mai de succes 200 de firme ale acestei țări.


TWO STEP CLUSTER
5 variabile continue și o variabilă categorială
"Log-likelihood" pentru a măsura distanța
Criteriul de clusterizare Bayesian (BIC)

# Tabel 1

Conform acestui tabel, cea mai bună soluție este cea care are cea mai mică valoare a criteriului Bayesian (de pe a 2-a coloană), precum și cea mai mare valoare de pe ultima coloană (Ratio of Distance Measures), așadar varianta cu 4 clustere.

*Table 1.* Automatic clustering

| Number of clusters | Schwarz's Bayesian Criterion (BIC) | BIC Change (a) | Ratio of BIC Changes (b) | Ratio of Distance Measures (c) |
|---|---|---|---|---|
| 1 | 1272,747 | | | |
| 2 | 990,722 | -282,026 | 1,000 | 1,839 |
| 3 | 884,535 | -106,186 | 0,377 | 1,152 |
| 4 | 806,071 | -78,464 | 0,278 | 2,168 |
| 5 | 825,571 | 19,500 | -0,069 | 1,906 |
| 6 | 884,966 | 59,395 | -0,211 | 1,212 |
| 7 | 952,054 | 67,088 | -0,238 | 1,394 |
| 8 | 1029,408 | 77,354 | -0,274 | 1,034 |
| 9 | 1107,607 | 78,199 | -0,277 | 1,146 |
| 10 | 1189,023 | 81,416 | -0,289 | 1,209 |
| 11 | 1274,238 | 85,216 | -0,302 | 1,015 |
| 12 | 1359,726 | 85,488 | -0,303 | 1,318 |
| 13 | 1449,533 | 89,808 | -0,318 | 1,373 |
| 14 | 1543,036 | 93,503 | -0,332 | 1,189 |
| 15 | 1638,112 | 95,076 | -0,337 | 1,126 |

(a) The changes are from the previous number of clusters in the table.
(b) The ratios of changes are relative to the change for the two cluster solution.
(c) The ratios of distance measures are based on the current number of clusters against the previous number of clusters.
*Source: Own creation*

# Tabel 2

Al doilea tabel (Cluster distribution) prezintă distribuția observațiilor în clustere, sau numărul de observații din fiecare cluster. De asemenea, tabelul cuprinde 24 de observații care sunt excluse, deoarece nu sunt suficient de relevante încât să fie grupate în clustere.

*Table 2.* Cluster distribution

| Cluster | Number of observations | % of combined cluster | % of total |
|---|---|---|---|
| 1 | 7 | 4,0% | 3,5% |
| 2 | 64 | 36,4% | 32,0% |
| 3 | 43 | 24,4% | 21,5% |
| 4 | 62 | 35,2% | 31,0% |
| Combined clusters | 176 | 100,0% | 88,0% |
| Excluded observations | 24 | | 12,0% |
| Total | 200 | | 100,0% |

*Source:* Own creation

# Tabel 3

Tabelul 3 (Cluster centroids) arată statisticile descriptive pentru variabilele continue. Valorile medii pentru toate variabilele continue din fiecare cluster sunt prezentate.

*Table 3.* Cluster centroids

| Variables | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| Total income in EUR (2007) | 301.044.003 | 25.714.168 | 21.771.629 | 16.452.301 |
| Total income in EUR (2006) | 248.912.033 | 21.215.261 | 16.912.291 | 13.322.724 |
| Income growth rate 2007/2006 | 27 | 33 | 6.119 | 129 |
| Earnings before tax EUR (2007) | 48.756.644 | 2.103.284 | 1.714.597 | 845.832 |
| Number of employees (2007) | 1.872 | 298 | 371 | 88 |

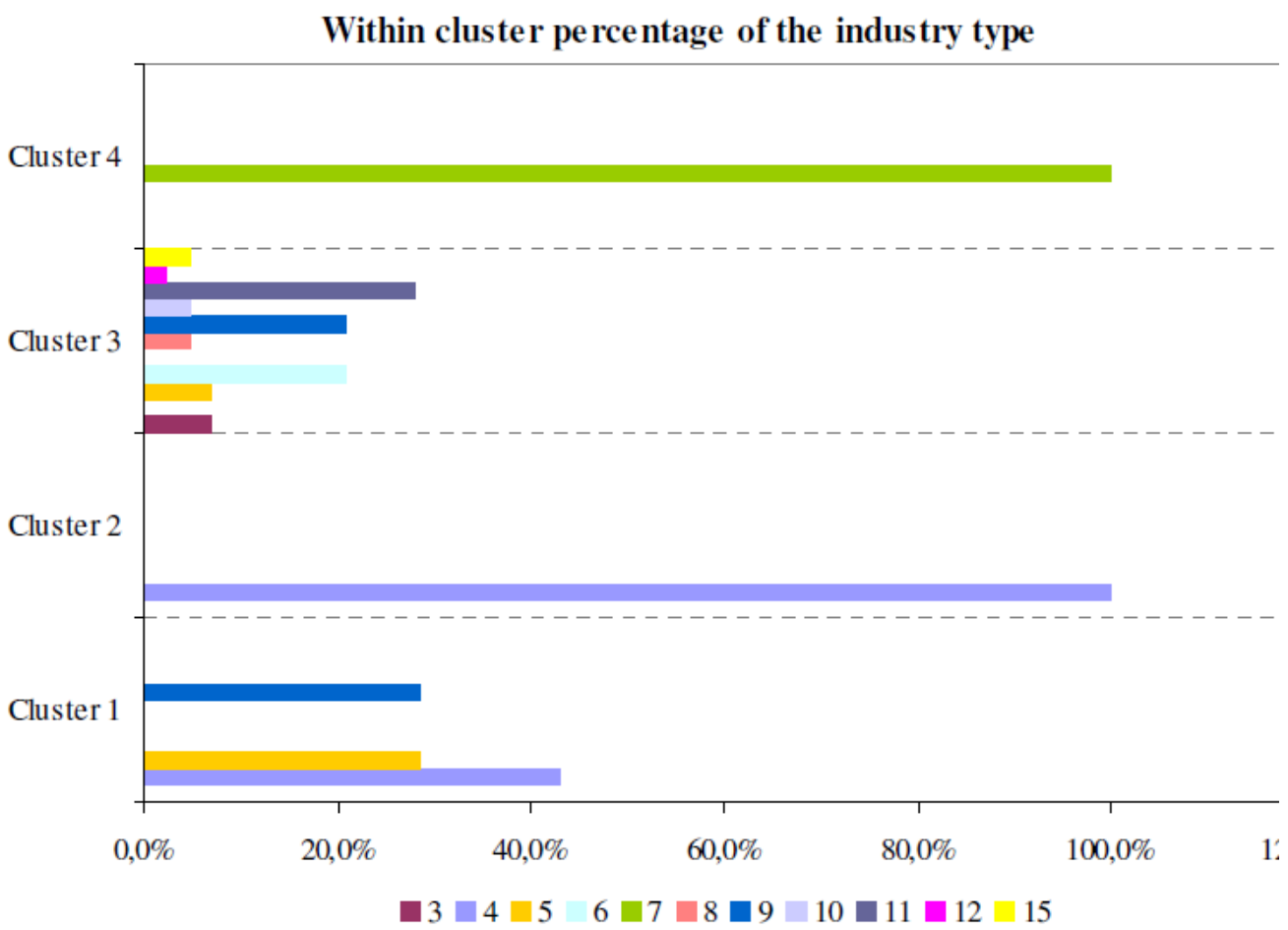
*Source:* Own creation

# Figura 1

- Figura 1 este o reprezentare grafică a frecvențelor celor 4 clustere (în procente).

- Acest grafic face referire la variabila categorială: tipul industriei.

*Figure 1.* Graphical presentation of the frequencies (in percentage)

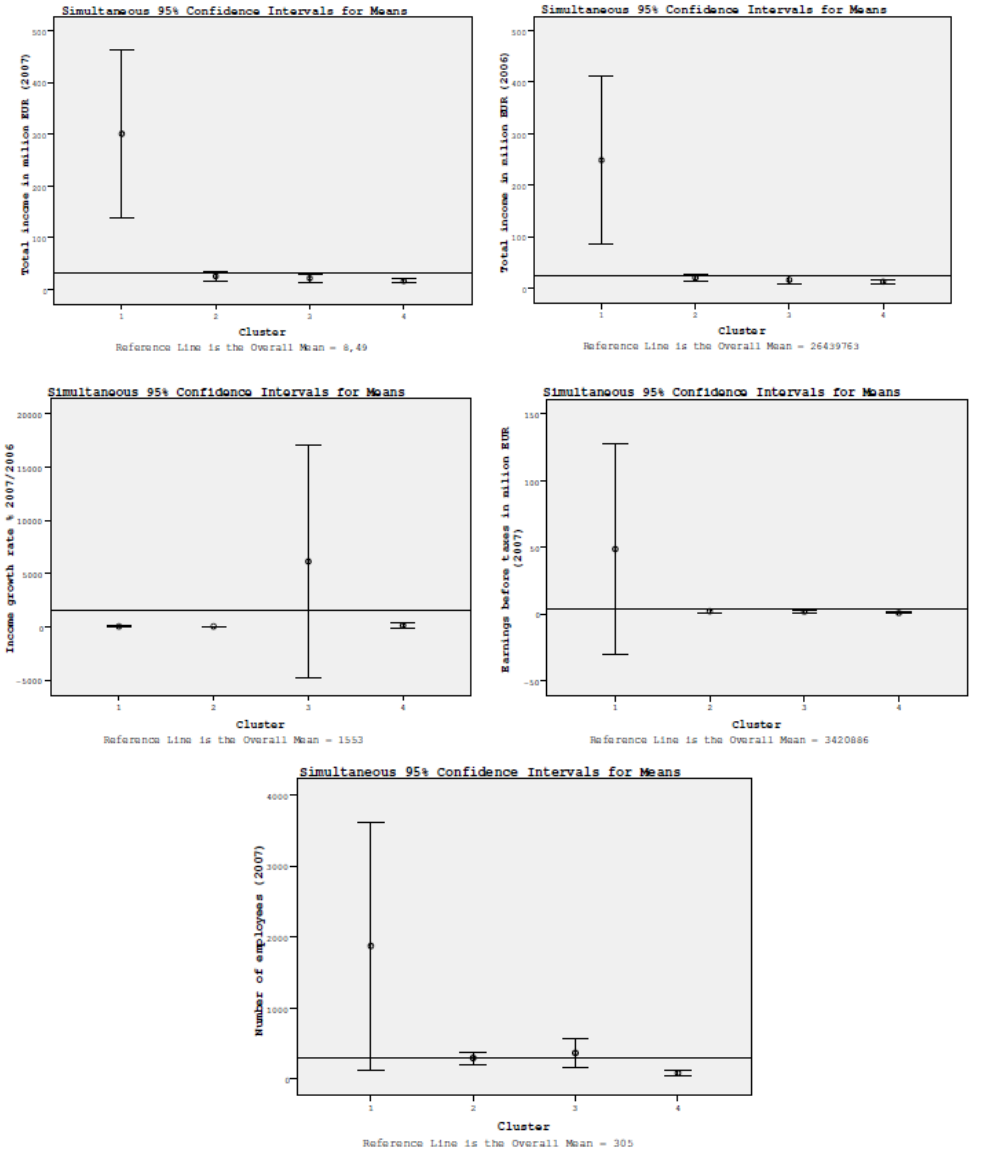



*Source:* Own creation

# Figura 2

Figura 2 ne arată variația din interiorul clusterelor pentru toate variabilele continue. S-a folosit un eșantion cu un interval de încredere de 95%. Primul boxplot reprezintă media variabilei „veniturile totale din anul 2007”, următoarele fiind pentru celelalte variabile scale.

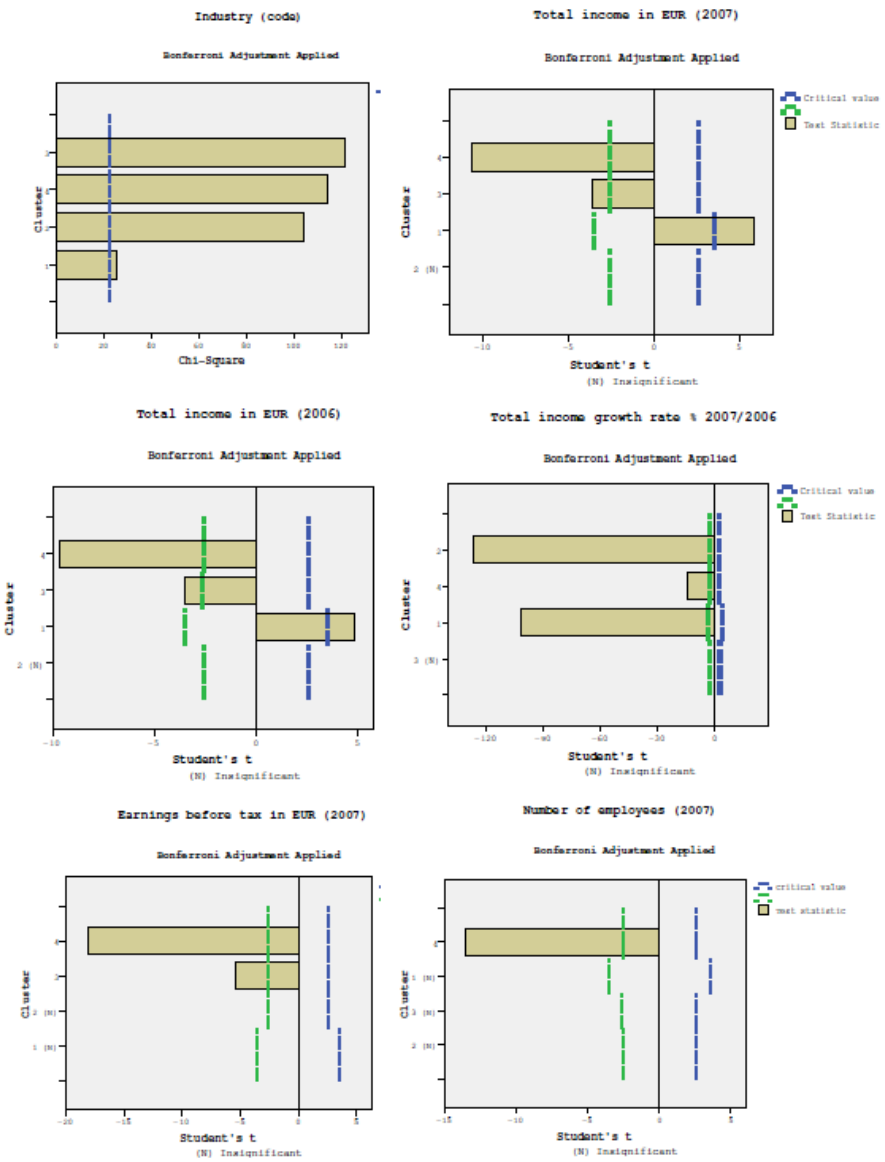
*Figure 2.* Within cluster variation for all continuous variables



# Figura 3

SPSS oferă încă un rezultat al analizei TwoStep, și anume grafice care arată semnificația variabilelor (Figura 3).

*Figure 3.* Variablewise importance




*Source:* Own creation

TWO STEP CLUSTER

# Concluzia

Conform rezultatelor empirice prezentate, analiza reușește să creeze soluții pentru 4 clustere sau 4 tipuri de companii diferite, fiind o bună reprezentare a pieței macedoniene.

TWO STEP CLUSTER

# Identificarea profilului clienților unei bănci

- Analiza datelor referitoare la împrumuturile realizate la o bancă din Germania.
- Eșantion: 1000 înregistrări
- 9 variabile nominale (categoriale) și 7 variabile continue
- Măsurarea distanței: log-likelihood
- Metoda Schwarz's Bayesian Criterion (BIC) pentru a determina numărul optim de clustere

TWO STEP CLUSTER

**Table 1.** Source data

| Duration | CreditHistory | Purpose | CreditAmount | YearsEmployed | PaymentRate | PersonalStatus |
|---|---|---|---|---|---|---|
| 6 | critical | television | 1169.0 | >=7 | 4.0 | male_single |
| 48 | ok_til_now | television | 5951.0 | <4 | 2.0 | female |
| 12 | critical | education | 2096.0 | <7 | 2.0 | male_single |
| 42 | ok_til_now | furniture | 7882.0 | <7 | 2.0 | male_single |
| 24 | past_delays | car_new | 4870.0 | <4 | 3.0 | male_single |
| 36 | ok_til_now | education | 9055.0 | <4 | 2.0 | male_single |
| 24 | ok_til_now | furniture | 2835.0 | >=7 | 3.0 | male_single |
| 36 | ok_til_now | car_used | 6948.0 | <4 | 2.0 | male_single |
| 12 | ok_til_now | television | 3059.0 | <7 | 2.0 | male_divorced |
| 30 | critical | car_new | 5234.0 | unemployed | 4.0 | male_married |

TWO STEP CLUSTER

# Auto-Clustering

**Table 2.** Auto-Clustering

| Number of clusters | Schwarz's Bayesian Criterion (BIC) | Ratio of Distance Measures |
|---|---|---|
| 1 | 26154,864 | |
| 2 | 25113,605 | 1,438 |
| 3 | 24488,152 | 1,542 |
| 4 | 24196,817 | 1,210 |
| 5 | 24012,292 | 1,189 |
| 6 | 23908,626 | 1,185 |
| 7 | 23871,920 | 1,131 |
| 8 | 23877,138 | 1,062 |
| 9 | 23900,958 | 1,009 |
| 10 | 23927,369 | |
| 11 | 23982,083 | 1,125 |
| 12 | 24066,712 | 1,047 |
| 13 | 24162,160 | 1,002 |
| 14 | 24258,157 | 1,030 |
| 15 | 24360,773 | 1,003 |

numărul optim de clustere este 3 deoarece numărul cel mai mare din coloana „Ratio of Distance Measures" se alfă pe al 3-lea rând.

TWO STEP CLUSTER

# Distribuția clusterelor

**Table 3.** Cluster distribution

| | N | % of Combined | % of Total |
|---|---|---|---|
| Cluster 1 | 150 | 15,0% | 15,0% |
| 2 | 351 | 35,1% | 35,1% |
| 3 | 499 | 49,9% | 49,9% |
| Combined | 1000 | 100% | 100% |
| Total | 1000 | | 100% |

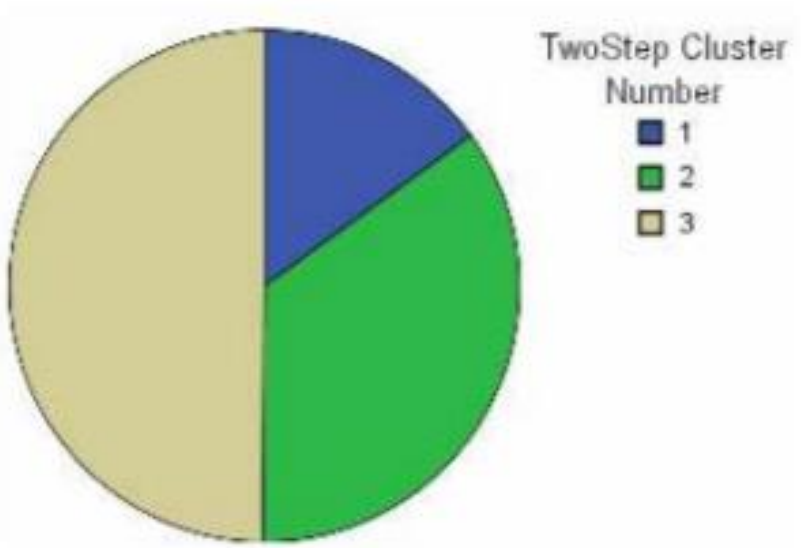



**Fig. 1.** Cluster size

TWO STEP CLUSTER

# Frecvențele variabilelor categoriale

**Table 4.** Frequencies for *SavingsAccount* variable

| | <100 | | <1000 | | <500 | | >=1000 | | unknown | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F | % | F | % | F | % | F | % | F | % |
| Cluster 1 | 88 | 14,6% | 8 | 12,7% | 16 | 15,5% | 5 | 10,4% | 33 | 18,0% |
| 2 | 240 | 39,8% | 20 | 31,7% | 34 | 33,0% | 15 | 31,3% | 42 | 23,0% |
| 3 | 275 | 45,6% | 35 | 55,6% | 53 | 51,5% | 28 | 58,3% | 108 | 59,0% |
| Combined | 603 | 100,0% | 63 | 100,0% | 103 | 100,0% | 48 | 100,0% | 183 | 100,0% |

*Note: * F –Frequencies; % - Percent*

TWO STEP CLUSTER

# Contribuția variabilei Property în fiecare cluster

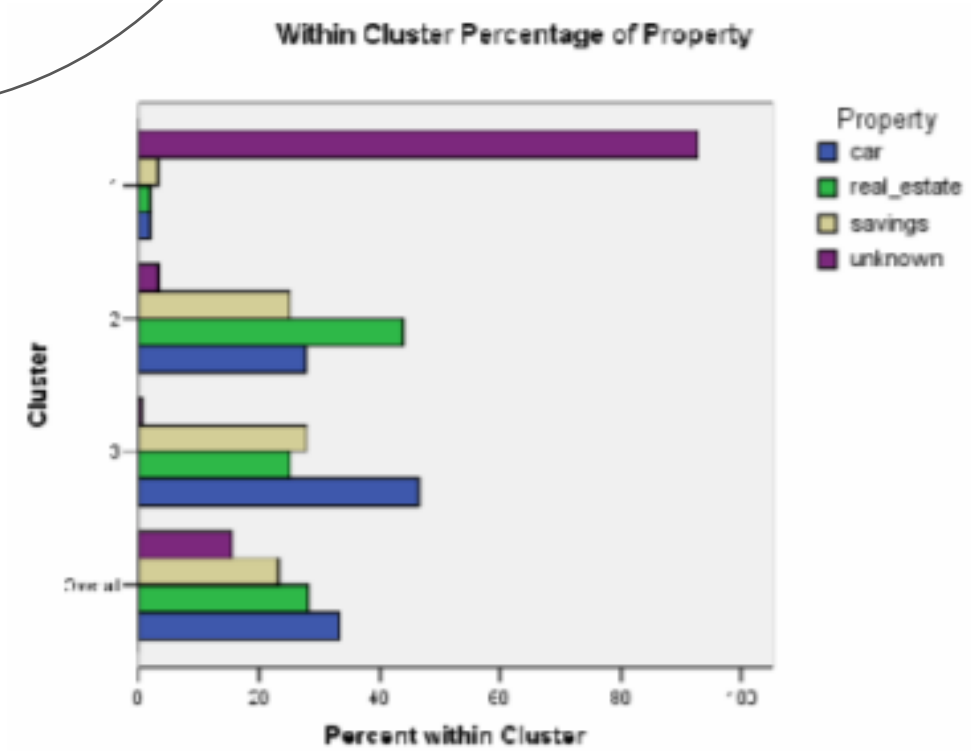



**Fig. 2.** The weight of *Property* in each cluster

În primul cluster proprietatea predominantă este necunoscută, în timp ce în clusterul 2 este proprietatea imobiliară și în cel de-al 3-lea este mașina.

TWO STEP CLUSTER

# Importanța variabilelor pentru fiecare cluster

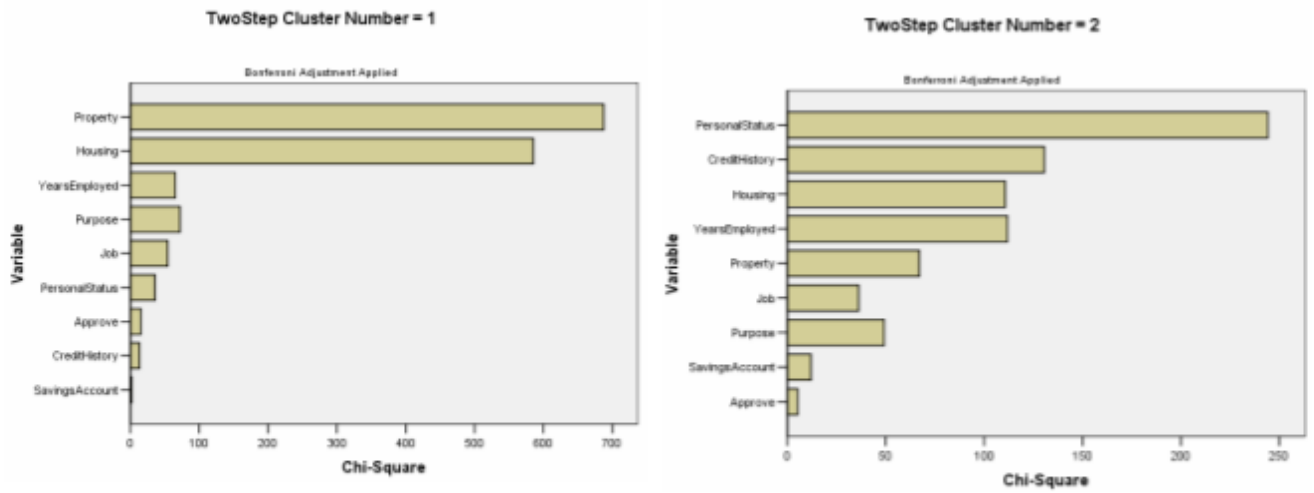



**Fig. 3.** Categorical variablewise importance for clusters 1 and 2

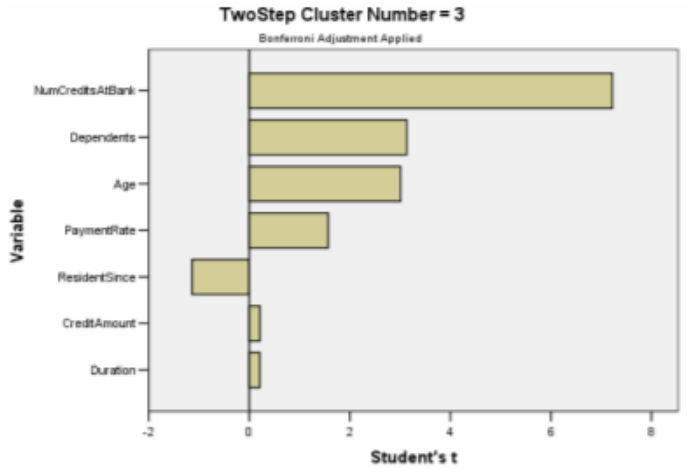



**Fig. 4.** Continuous variablewise importance for cluster 3

TWO STEP CLUSTER

# Concluzia

Folosind procedura TwoStep Cluster Analysis, s-au creat 3 profiluri de clienți.

- Clienți calificați
  Educație
  afaceri
- Imobiliare
  Șomeri
  Recalificare
  Articole menaj
- Mașina
  Televizor
  educație

# Mulțumim!