Proiect la Analiza Datelor

Masterand: Darie Maxim

Specializarea: TAPI

Grupa: 41307

An: I

Cuprins

Car Evaluation Data Set

Formularea problemei

Baza de date auto de evaluare a fost derivată dintr-un simplu model de decizie ierarhic dezvoltat inițial pentru demonstrarea DEX. Analiza statistica a datelor auto a fost folosita pentru descoperirea tiparelor si pentru a explica diferentele dintre subseturi de date.

In calitate de constructor in automobilistica pentru a produce modele de autovehicule cat mai economе din punct de vedere a consumului de carburant(nmg) am avut necesitatea de a manipula setul de date “Car Evaluation Data Set” pentru a pruduce un vehicul cat mai economic.

Metode și tehnici de soluționare ale problemei

Pentru a reduce esential consumul de combustibil este necesar sa lual in calcul asa factori ca :

1.Ajustarea mărimii vehiculelor depinde de tipul activității desfășurate cu ajutorul lor. E decizia managementului dacă se poate opta pentru mașinii mai mici cu consum mai bun de combustibil la 100 km.

2. Tot mai mulți manageri doresc să includă în flota companiei mașini hibrid, care folosesc energie verde, în special pentru activitatea desfășurată în oraș. Mașinile

hibrid sunt mai eficiente și mai puțin poluante, deci reduc emisiile și cresc economiile de carburant, cu efecte pozitive asupra mediului și bugetului companiei.

Experimente efectuate

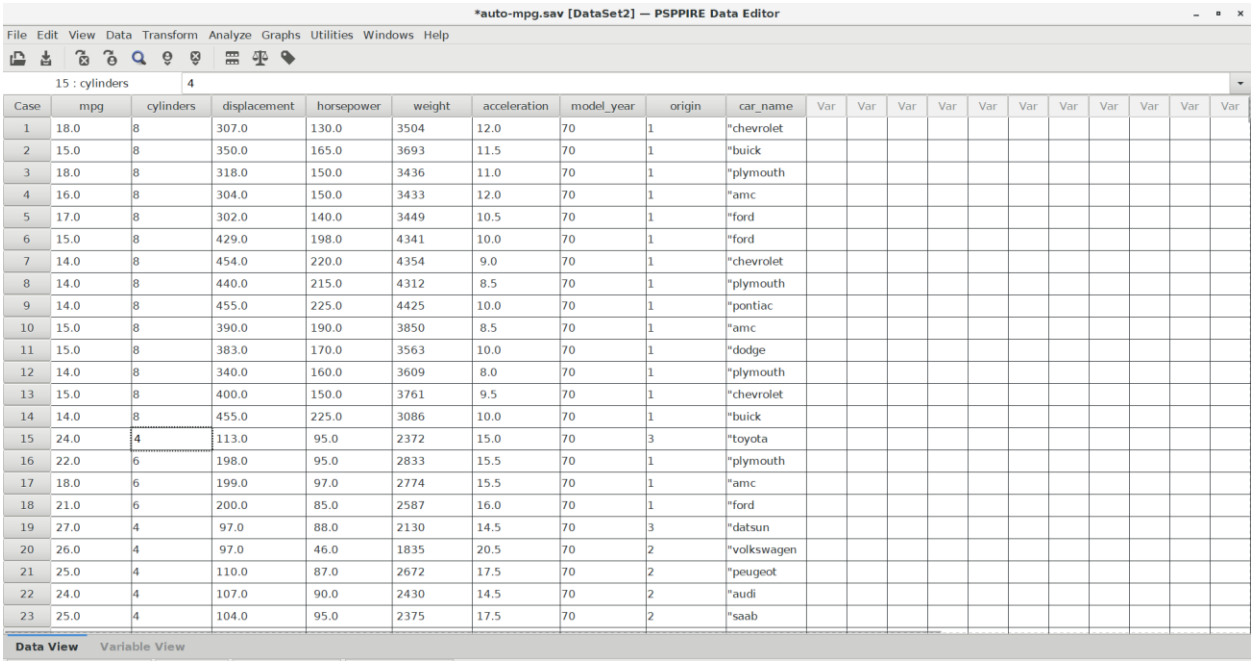
Primul pas efectuat pentru analiza Car Evaluation Data Set a fost download-area fisierelor sursa Data Folder/Data SetDescription din baza de date publică UCI Machine Learning Repository.

Derivată dintr-un model de decizie ierarhic simplu, această bază de date poate fi utilă pentru testarea inducției constructive și a metodelor de descoperire a structurii.

Fisierele sursa download-ate (car.c45-names, car.dat, car.names) inainte de a face orice fel de analiză statistică cu PSPP, au fost redactate cu ajutorul aplicatiei Notepad++

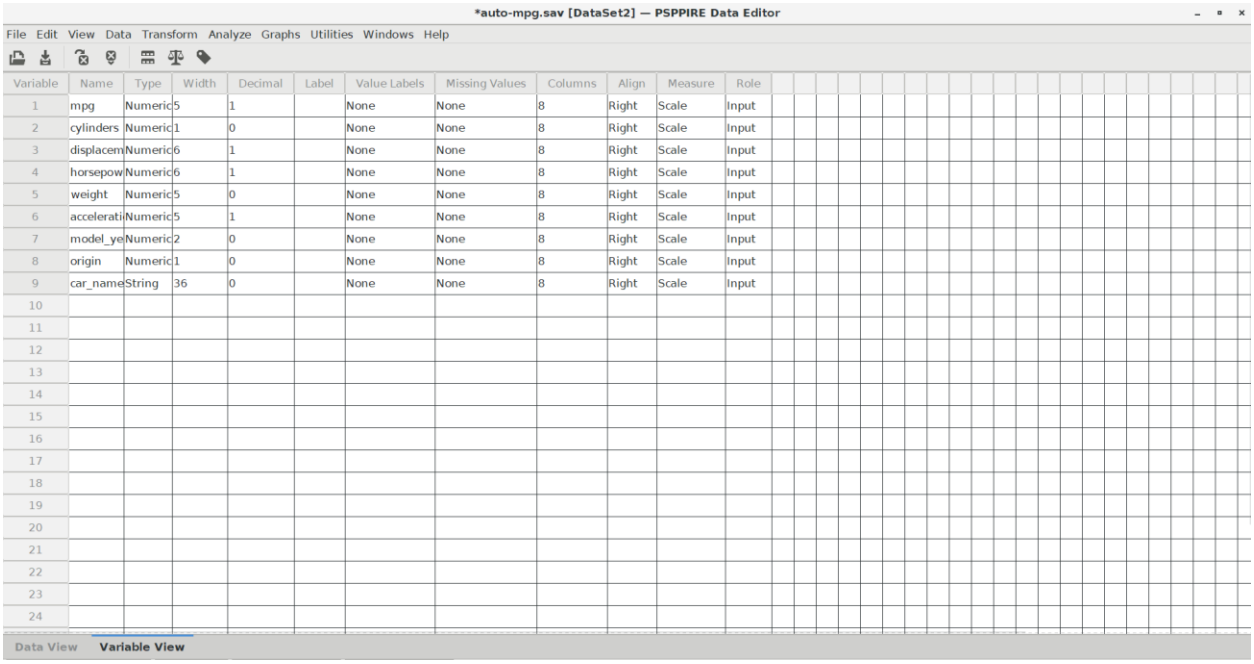
Din motiv ca a fost cu valori lipsă, a fost redactata si modalitatea de aliniere , doar apoi salvate ca fisiere cu extensia .sav si apoi importante in aplicatia grafica PSPP.

DATA VIEW:



Case	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	car_name
1	18.0	8	307.0	130.0	3504	12.0	70	1	"chevrolet
2	15.0	8	350.0	165.0	3693	11.5	70	1	"buick
3	18.0	8	318.0	150.0	3436	11.0	70	1	"plymouth
4	16.0	8	304.0	150.0	3433	12.0	70	1	"amc
5	17.0	8	302.0	140.0	3449	10.5	70	1	"ford
6	15.0	8	429.0	198.0	4341	10.0	70	1	"ford
7	14.0	8	454.0	220.0	4354	9.0	70	1	"chevrolet
8	14.0	8	440.0	215.0	4312	8.5	70	1	"plymouth
9	14.0	8	455.0	225.0	4425	10.0	70	1	"pontiac
10	15.0	8	390.0	190.0	3850	8.5	70	1	"amc
11	15.0	8	383.0	170.0	3563	10.0	70	1	"dodge
12	14.0	8	340.0	160.0	3609	8.0	70	1	"plymouth
13	15.0	8	400.0	150.0	3761	9.5	70	1	"chevrolet
14	14.0	8	455.0	225.0	3086	10.0	70	1	"buick
15	24.0	4	113.0	95.0	2372	15.0	70	3	"toyota
16	22.0	6	198.0	95.0	2833	15.5	70	1	"plymouth
17	18.0	6	199.0	97.0	2774	15.5	70	1	"amc
18	21.0	6	200.0	85.0	2587	16.0	70	1	"ford
19	27.0	4	97.0	88.0	2130	14.5	70	3	"datsun
20	26.0	4	97.0	46.0	1835	20.5	70	2	"volkswagen
21	25.0	4	110.0	87.0	2672	17.5	70	2	"peugeot
22	24.0	4	107.0	90.0	2430	14.5	70	2	"audi
23	25.0	4	104.0	95.0	2375	17.5	70	2	"saab

VARIABLE VIEW:



Variable	Name	Type	Width	Decimal	Label	Value Labels	Missing Values	Columns	Align	Measure	Role
1	mpg	Numeric	5	1		None	None	8	Right	Scale	Input
2	cylinders	Numeric	1	0		None	None	8	Right	Scale	Input
3	displacem	Numeric	6	1		None	None	8	Right	Scale	Input
4	horsepow	Numeric	6	1		None	None	8	Right	Scale	Input
5	weight	Numeric	5	0		None	None	8	Right	Scale	Input
6	accelerati	Numeric	5	1		None	None	8	Right	Scale	Input
7	model_ye	Numeric	2	0		None	None	8	Right	Scale	Input
8	origin	Numeric	1	0		None	None	8	Right	Scale	Input
9	car_name	String	36	0		None	None	8	Right	Scale	Input

Analiza exploratorie a datelor

Pentru a calcula valoarea minimă, valoarea maximă, media,

abaterea standard, asimetria pentru variabila (cylinders) a fost folosit Descriptive Statistics Frequencies.

Valid cases = 398; cases with missing value(s) = 6.

Variable	*N*	*Mean*	*Std Dev*	*Minimum*	*Maximum*
mpg	398	23.51	7.82	9.00	46.60
cylinders	398	5.45	1.70	3.00	8.00
displacement	398	193.43	104.27	68.00	455.00
horsepower	392	104.47	38.49	46.00	230.00
weight	398	2970.42	846.84	1613.00	5140.00
acceleration	398	15.57	2.76	8.00	24.80
model_year	398	76.01	3.70	70.00	82.00
origin	398	1.57	.80	1.00	3.00

```
FREQUENCIES

FREQUENCIES
      /VARIABLES= cylinders
      /FORMAT=AVALUE TABLE.
```

cylinders

Value Label	*Value*	*Frequency*	*Percent*	*Valid Percent*	*Cum Percent*
	3	4	1.01	1.01	1.01
	4	204	51.26	51.26	52.26
	5	3	.75	.75	53.02
	6	84	21.11	21.11	74.12
	8	103	25.88	25.88	100.00
	Total	398	100.0	100.0	

cylinders

N	*Valid*	398
	Missing	0
Mean		5.45
Std Dev		1.70
Minimum		3.00
Maximum		8.00

În ambele tabele se regăsesc valorile variabilei (cylinders) , frecvențele absolute, procentul, procentul

cumulat, în timp ce al doilea tabel conține valorile indicatorilor statistici medie, abaterea

standard, asimetria, valoarea maximă, respectiv valoarea minimă.

Calcularea statisticilor descriptive: valoarea minimă, valoarea maximă, media, abaterea

standard, dispersia și asimetria pentru atributul (WEIGHT).

```
DESCRIPTIVES
  /VARIABLES= weight
  /STATISTICS=DEFAULT VARIANCE SKEWNESS.
```

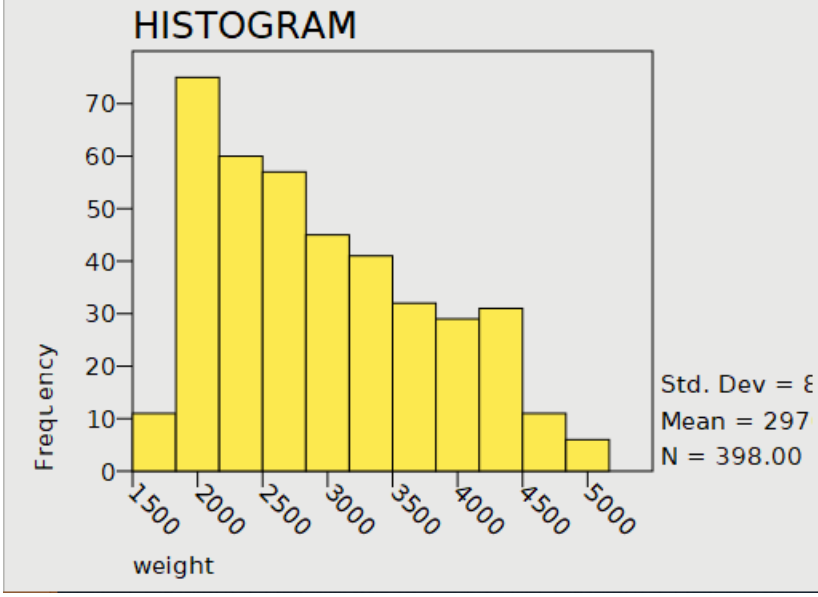
Valid cases = 398; cases with missing value(s) = 0.

Variable	*N*	*Mean*	*Std Dev*	*Variance*	*Skewness*	*S.E. Skew*	*Minimum*	*Maximum*
weight	398	2970.42	846.84	717140.99	.53	.12	1613.00	5140.00

Reprezentarea grafic pentru variabila (WEIGHT), utilizând histograma:

GRAPH

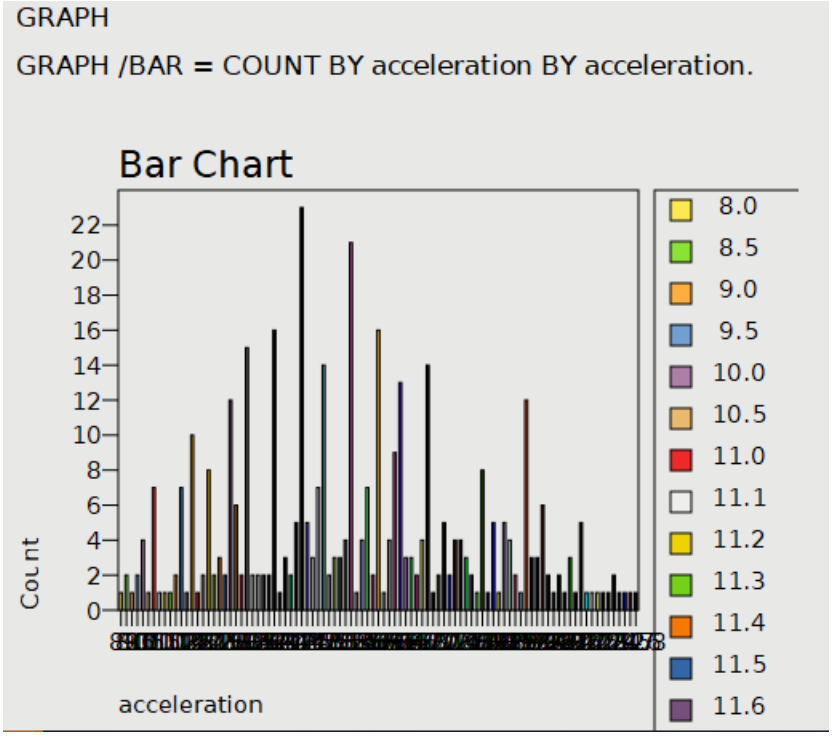
GRAPH /HISTOGRAM = weight.



Reprezentarea grafică sub fomă de bar chart pentru variabila (ACCELERATION):

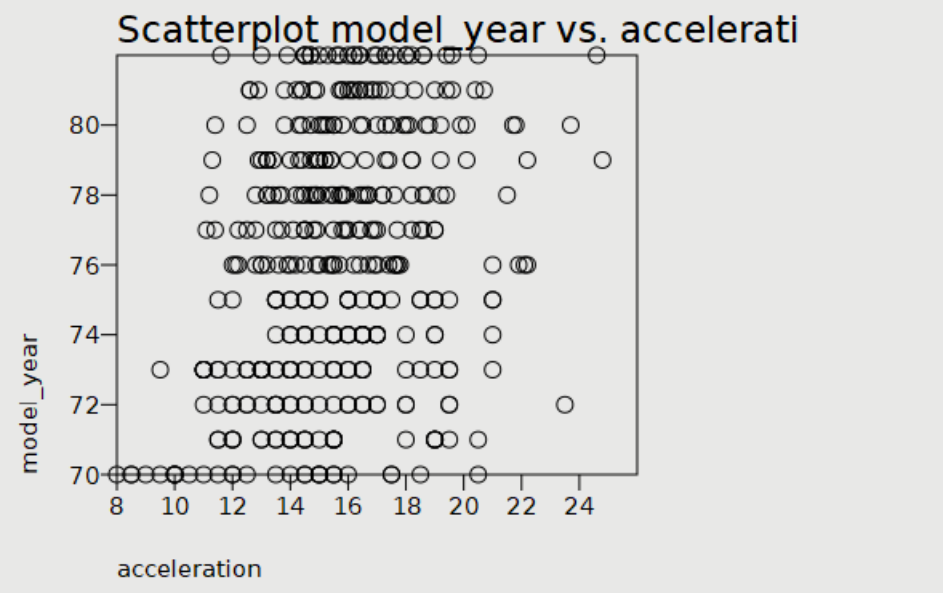
GRAPH

GRAPH /BAR = COUNT BY acceleration BY acceleration.



Am realizat un grafic de tip scatter plot și am interpretat rezultatele. Am droit sa verific daca există o legătură între anul de fabricatie a autoturismului (variabila: model_year) si aceleratie (variabila: acceleration)

```
GRAPH
GRAPH SCATTERPLOT(BIVARIATE) = acceleration WITH model_year.
```



Dupa care a fost analizata corelația dintre variabila(displacement) ce reprezinta distanta parcursa si (mpg) care reprezinta consumul de combustibil . Astfel, între distanta de deplasare și consumul de carburant 1L/100km există o corelație de -0.80 ceaia ce semnifica o corelatie puternica negative. Variabilele sunt invers proportionale.

```
CORRELATION
      /VARIABLES =  mpg displacement
      /PRINT = TWOTAIL NOSIG.
```

Correlations

		mpg	displacement
mpg	Pearson Correlation	1.00	-.80
	Sig. (2-tailed)		.000
	N	398	398
displacement	Pearson Correlation	-.80	1.00
	Sig. (2-tailed)	.000	
	N	398	398

Pentru a furniza reprezentări grafice de date, cum ar fi histograme sau boxplot-uri se consideră fișierul sursă auto-mpg.sav.

Corelația variabilelor

```
CORRELATIONS
CORRELATION
      /VARIABLES =  acceleration model_year mpg cylinders displacement horsepower weight origin
      /PRINT = TWOTAIL SIG.
```

Correlations

		acceleration	model_year	mpg	cylinders	displacement	horsepower	weight	origin
acceleration	Pearson Correlation	1.00	.29	.42	-.51	-.54	-.69	-.42	.21
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000
	N	398	398	398	398	398	392	398	398
model_year	Pearson Correlation	.29	1.00	.58	-.35	-.37	-.42	-.31	.18
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000
	N	398	398	398	398	398	392	398	398
mpg	Pearson Correlation	.42	.58	1.00	-.78	-.80	-.78	-.83	.56
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000
	N	398	398	398	398	398	392	398	398
cylinders	Pearson Correlation	-.51	-.35	-.78	1.00	.95	.84	.90	-.56
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000
	N	398	398	398	398	398	392	398	398
displacement	Pearson Correlation	-.54	-.37	-.80	.95	1.00	.90	.93	-.61
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000
	N	398	398	398	398	398	392	398	398
horsepower	Pearson Correlation	-.69	-.42	-.78	.84	.90	1.00	.86	-.46
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000
	N	392	392	392	392	392	392	392	392
weight	Pearson Correlation	-.42	-.31	-.83	.90	.93	.86	1.00	-.58
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000
	N	398	398	398	398	398	392	398	398
origin	Pearson Correlation	.21	.18	.56	-.56	-.61	-.46	-.58	1.00
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	
	N	398	398	398	398	398	392	398	398

Se dorește a se verifica ipoteza următoare: Greutatea media a vehiculelor din lista este de 2600 kg (H0=2600). Intervalul de încredere considerat este 95%(default). Astfel, pentru ipoteza enunțată, se alege variabila (weight) și se precizează valoarea medie cu care se va compara media variabilei WEIGHT (H0=50).

```
T-TEST
T-TEST /TESTVAL=2600
     /VARIABLES= weight    /MISSING=ANALYSIS
     /CRITERIA=CI(0.95).
```

One-Sample Statistics

	N	Mean	Std. Deviation	S.E. Mean
weight	398	2970.42	846.84	42.45

One-Sample Test

	Test Value = 2600.000000					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
weight	8.73	397	.000	370.42	286.97	453.88

Rezultatele obținute indică o valoare p=0.000 și o valoare t=8.73 si df=397.

ANOVA

Deoarece variabila categorială *cylinders* din fișierul sursă are mai mult de două valori, trebuie să folosim modelul ANOVA unifactorial .Presupunem situatia: dorim sa cunoastem daca numarul de cilindri la masina afecteaza acceleratia,

Analyze → Compare Means → One Way ANOVA

primul tabel reprezintă statisticile descriptive (mediile pentru fiecare grup în parte, abaterea standard, eroarea standard, coeficientul de încredere (limita inferioară și limita superioară), minimul și maximul; al doilea tabel se referă la testul de omogenitate a variației și conține *testul Levene*, gradele de libertate *df1, df2 și* semnificația; al treilea tabel rezumă testul ANOVA și conține variația inter

grupuri (SSA), variația intra grupuri (SSW) și variația totală (SST), testul F și semnificația.

Nota: Pe exemplu de date care a fost selectat nu pot efectua Testul ANOVA

Voi folosi matricea de covariatie deoarece valorile nu sunt normalizate(While correlation coefficients lie between -1 and +1)

Communalities

	Initial	Extraction
cylinders	2.90	2.63
displacement	10922.43	10922.43
horsepower	1477.79	1477.79
weight	719644.19	719644.19
acceleration	7.59	7.59
model_year	13.54	13.54
origin	.65	.28

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	730283.94	99.76	99.76	730283.94	99.76	99.76	589335.53	80.50	80.50
2	1510.11	.21	99.96	1510.11	.21	99.96	47716.51	6.52	87.02
3	260.58	.04	100.00	260.58	.04	100.00	59588.75	8.14	95.16
4	10.94	.00	100.00	10.94	.00	100.00	11502.18	1.57	96.73
5	2.87	.00	100.00	2.87	.00	100.00	23925.47	3.27	100.00
6	.38	.00	100.00						
7	.26	.00	100.00						

Component Matrix

	Component				
	1	2	3	4	5
cylinders	1.53	.52	-.12	.00	-.03
displacement	97.71	36.75	-4.90	-.03	.02
horsepower	33.30	11.59	15.31	-.14	.15
weight	848.31	-4.69	-.04	.00	-.01
acceleration	-1.16	-1.35	-1.24	.20	1.69
model_year	-1.14	-.93	-.69	-3.30	.09
origin	-.47	-.13	.20	-.01	.00

Rotated Component Matrix

	Component				
	1	2	3	4	5
cylinders	1.50	-.06	.29	.27	.48
displacement	95.09	-7.27	21.20	18.15	32.38
horsepower	25.53	.83	20.23	8.43	18.58
weight	761.34	218.32	242.34	105.30	150.08
acceleration	-.63	-.02	-.24	-.35	-2.65
model_year	-.60	-.02	-.22	-3.59	-.47
origin	-.52	.02	.03	-.06	-.05

Bibliografie

1. https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

S-a parcurs întreaga arie a cerințelor iar analiza datelor a fost una amănunțită și documentată.