Proiect la Analiza Datelor

Masterand: Darie Maxim

Specializarea: TAPI

Grupa: 41307

An: I

Cuprins

Formularea problemei .................................................................................................. 3

Metode și tehnici de soluționare ale problemei ........................................................... 3

Experimente efectuate .................................................................................................. 3

Încărcarea fișierului sursă ......................................................................................... 4

Analiza exploratorie a datelor ................................................................................... 4

Corelația variabilelor ................................................................................................. 9

Teste statistice pentru compararea mediilor .......................................................... 10

Anova ....................................................................................................................... 11

Bibliografie .................................................................................................................. 11

**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 econome 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++

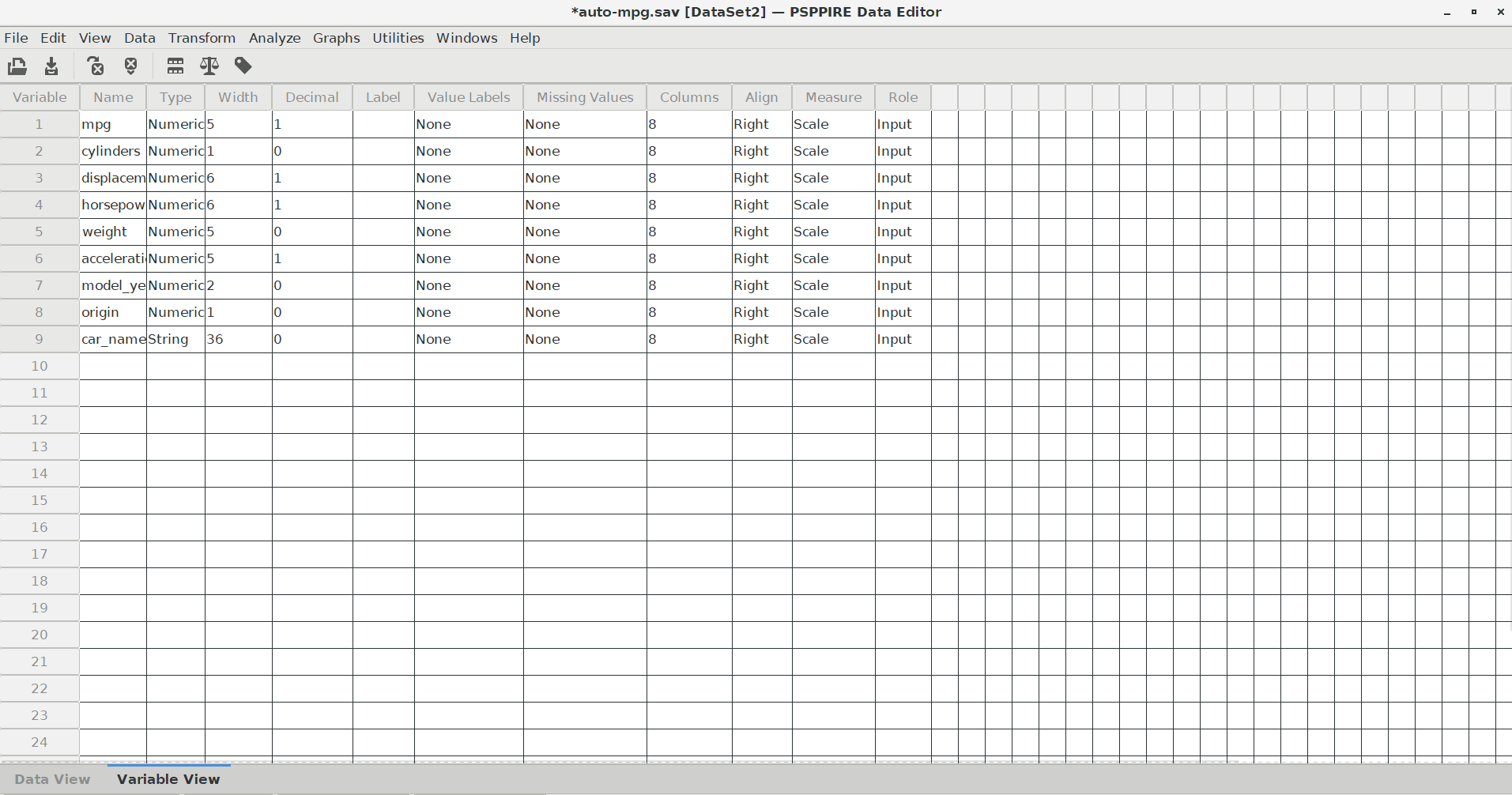
Din motiv ca a fost cu valori lipsă, a fost redactata si modalitatea de aliniere , doar apoi salvate ca fisiere cu extensia .sav si apoi importante in aplicatia grafica PSPP.

DATA VIEW:

A screenshot of a computer

Description automatically generated

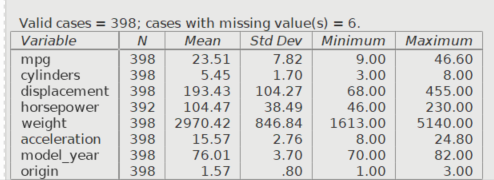
VARIABLE VIEW:



# Analiza exploratorie a datelor

Pentru a calcula valoarea minimă, valoarea maximă, media,

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



A screenshot of a cell phone

Description automatically generated

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).

A screenshot of a cell phone

Description automatically generated

Reprezentarea grafic pentru variabila (WEIGHT), utilizând histograma:

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

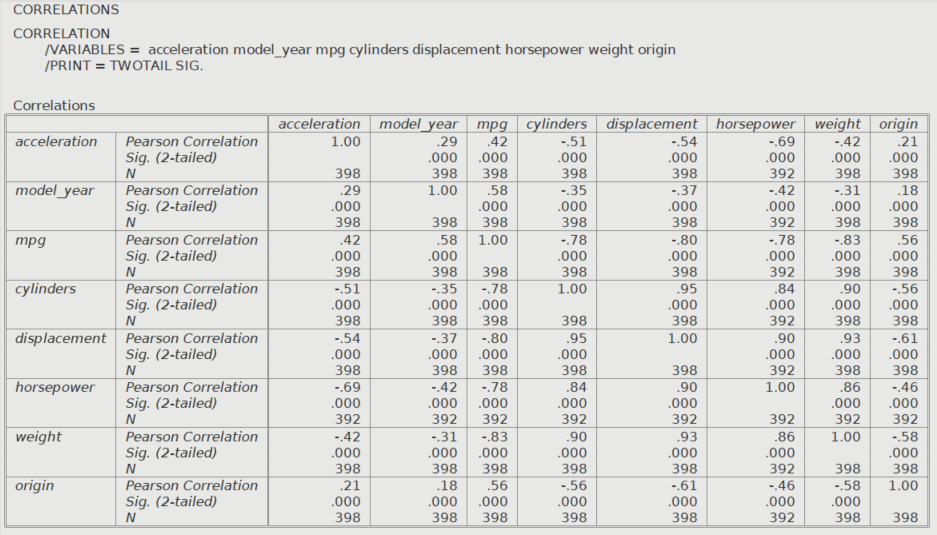
Dupa care a fost analizata corelația dintre variabila(displacement) ce reprezinta distanta parcursa si (mpg) care reprezinta consumul de combustibil . Astfel, între distanta de deplasare și consumul de carburant 1L/100km există o corelație de -0.80 ceaia ce semnifica o corelatie puternica negative. Variabilele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

A screenshot of a cell phone

Description automatically generated

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

Corelația variabilelor



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).

A screenshot of a cell phone

Description automatically generated

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

# ANOVA

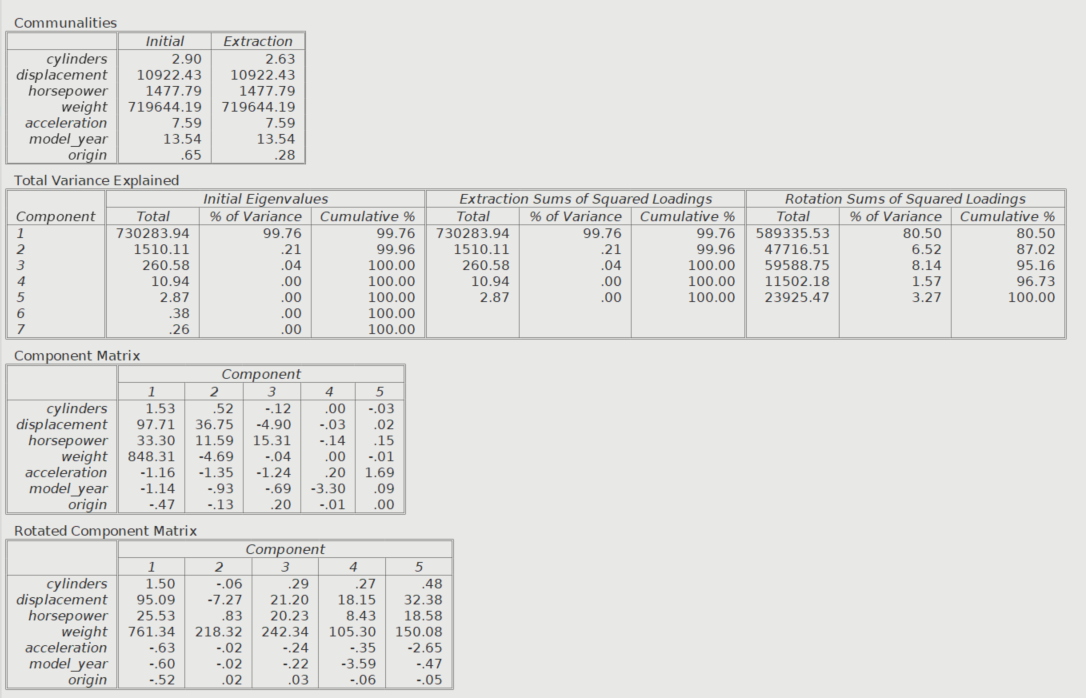
Deoarece variabila categorială *cylinders* din fișierul sursă are mai mult de două valori, trebuie să folosim modelul ANOVA unifactorial .Presupunem situatia: dorim sa cunoastem daca numarul de cilindri la masina afecteaza acceleratia,

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)



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.