Analiza asupra Autovehiculelor Electrice

Kutkut Sami, Luca Andreea Giorgiana, Mertic Yulyssa Nicoleta Ioana

22/01/2022

Table of Contents

[I. Prezentarea Domeniului 1](#_Toc125886771)

[II. Obiective 1](#_Toc125886772)

[III. Literatură de specialitate 1](#_Toc125886773)

[IV. Baza de date 2](#_Toc125886774)

[V. Analiza Descriptivă 3](#_Toc125886775)

[VI. Analiza Componentelor Principale 7](#_Toc125886776)

[VII. Arbori de clasificare 10](#_Toc125886777)

[VIII. Concluzii 12](#_Toc125886778)

## I. Prezentarea Domeniului

Prezenta analiză se concentrează pe observarea caracteristicilor autovehiculelor electrice dintr-un set de date preluat de pe site-ul Kaggel: <https://www.kaggle.com/code/prateekmaj21/electric-cars-data-analysis/data>  
În acest sens se dorește observarea diferențelor, asemănărilor si performanțelor setului înregistrat din punct de vedere al capacității bateriei, consumului de electricitate, distanța ce poate fi parcursă, viteza, preț si desigur a altor caracteristici.

## II. Obiective

Scopul analizei are ca obiectiv determinarea performanțelor principale ale mașinilor electrice utilizând metoda anlizei componentelor principale.  
In general, alegerea unui autovehicul este influentat de caracteristicile acestuia, prin urmare am ales sa realizam arbori de decizie si de clasificare cu ajutorul carora s-a putut contura o viziune mai clara asupra gruparilor obtinute.  
De asemenea, am realizat o clasificare supervizata pentru numarul de viteze prin algoritmul specific machine-learning, KNN.

## III. Literatură de specialitate

Vehiculele electrice (EV) reprezintă o tehnologie promițătoare pentru realizarea unui sector de transport durabil în viitor, datorită emisiilor de carbon foarte scăzute până la zero, zgomotului redus, eficienței ridicate și flexibilității în funcționarea și integrarea în rețea. Acest capitol include o prezentare generală a tehnologiilor vehiculelor electrice, precum și a sistemelor asociate de stocare a energiei și a mecanismelor de încărcare. Sunt prezentate diferite tipuri de vehicule cu propulsie electrică.

* " Nanaki, Evanthia A. (2021). Electric Vehicles for Smart Cities || Electric vehicles. , (), 13–49. <doi:10.1016/B978-0-12-815801-2.00006-X> " <https://www.sciencedirect.com/science/article/pii/B978012815801200006X>
* "Kittner, Noah (2020). *Technological Learning in the Transition to a Low-Carbon Energy System || Electric vehicles. , (), 145–163.* <doi:10.1016/B978-0-12-818762-3.00009-1> "  
  <https://www.sciencedirect.com/science/article/pii/B9780128187623000091>

## IV. Baza de date

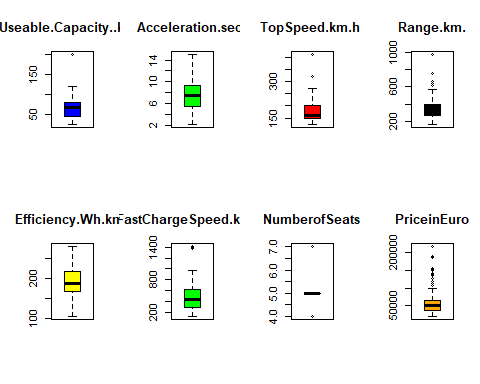
## # A tibble: 164 × 9  
## Name Useab…¹ Accel…² TopSp…³ Range…⁴ Effic…⁵ FastC…⁶ Numbe…⁷ Price…⁸  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Opel Ampera-e 58 7.3 150 335 173 210 5 42990  
## 2 Nissan Leaf 36 7.9 144 220 164 230 5 29990  
## 3 Porsche Tayc… 83.7 2.8 260 390 215 860 4 186336  
## 4 Nissan e-NV2… 36 14 123 165 218 170 7 43433  
## 5 Volkswagen I… 45 8.9 160 275 164 260 5 31960  
## 6 BMW iX3 74 6.8 180 385 192 520 5 66300  
## 7 Nissan Leaf … 56 7.3 157 325 172 390 5 38350  
## 8 BMW i3 120 Ah 37.9 7.3 150 235 161 270 4 39000  
## 9 Mercedes EQA… 66.5 8.9 160 355 187 420 5 47541  
## 10 DS 3 Crossba… 45 8.7 150 250 180 330 5 30040  
## # … with 154 more rows, and abbreviated variable names  
## # ¹​`Useable Capacity: kWh`, ²​`Acceleration(sec)`, ³​`TopSpeed(km/h)`,  
## # ⁴​`Range(km)`, ⁵​`Efficiency(Wh/km)`, ⁶​`FastChargeSpeed(km/h)`,  
## # ⁷​NumberofSeats, ⁸​PriceinEuro

* **Useable Capacity: kWh**  
  Capacitate utilizabilă sau „kilowați-oră (kWh) disponibile” reprezintă de obicei 95%-99% din capacitatea totală disponibilă. De exemplu, o baterie care poate ține fizic un total de 65 kWh poate pune la dispoziție numai 62 kWh pentru mașină.
* **Acceleration(sec)**  
  Acest indicator se referă la timpul necesar pentru a ajunge la 100 km/h. Timpul este exprimat în secunde.
* **TopSpeed(km/h)**  
  Viteza maximă la care poate ajunge mașina.
* **Range(km)**  
  Cât poate parcurge mașina cu un plin al bateriei.
* **Efficiency(Wh/km)** Eficiența este măsurată în wați-oră pe km, sau mai frecvent scrisă ca wh/km. Aceasta înseamnă că pentru fiecare kilometru parcurs, vehiculul dumneavoastră electric va consuma, în medie, xx wați oră din acumulator.
* **FastChargeSpeed(km/h)**  
  Măsoară câți km poate parcurge mașina cu o oră de încărcare în modul FastCharge.
* **NumberofSeats**  
  Numărul de locuri disponibil în mașină.
* **PriceinEuro**  
  Prețul în euro.

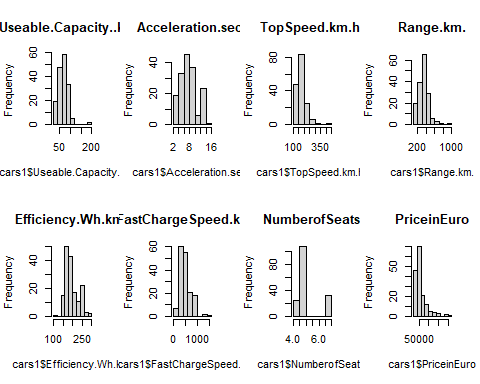
## V. Analiza Descriptivă

Pentru a analiza setul de date, am verificat existenta outlierilor. In acest sens, boxploturile de mai jos au evidentiat ca exista un numar mic de outlieri, neseminifcativi pentru atingerea obiectiviului analize.

cars1<-data.frame(Cars[,-1])  
#cars1  
par(mfrow=c(2,4))  
boxplot(cars1$Useable.Capacity..kWh, main="Useable.Capacity..kWh", col="blue")#Are un(1) outlier  
boxplot(cars1$Acceleration.sec., main="Acceleration.sec.", col="green") #Nu are outlieri  
boxplot(cars1$TopSpeed.km.h., main="TopSpeed.km.h.", col="red") #Are 2 outlieri  
boxplot(cars1$Range.km., main="Range.km.", col="black") #Are 5 outlieri  
boxplot(cars1$Efficiency.Wh.km., main="Efficiency.Wh.km.", col="yellow") #Nu are outlieri  
boxplot(cars1$FastChargeSpeed.km.h., main="FastChargeSpeed.km.h.", col="green") #Are 3-4outlieri  
boxplot(cars1$NumberofSeats, main="NumberofSeats", col="pink") #Are 2 outlieri  
boxplot(cars1$PriceinEuro, main="PriceinEuro", col="orange") #Are multi outlieri



par(mfrow=c(2,4))  
hist(cars1$Useable.Capacity..kWh, main="Useable.Capacity..kWh")  
hist(cars1$Acceleration.sec., main="Acceleration.sec.")  
hist(cars1$TopSpeed.km.h., main="TopSpeed.km.h.")  
hist(cars1$Range.km., main="Range.km.")  
hist(cars1$Efficiency.Wh.km., main="Efficiency.Wh.km.")  
hist(cars1$FastChargeSpeed.km.h., main="FastChargeSpeed.km.h.")  
hist(cars1$NumberofSeats, main="NumberofSeats")  
hist(cars1$PriceinEuro, main="PriceinEuro")

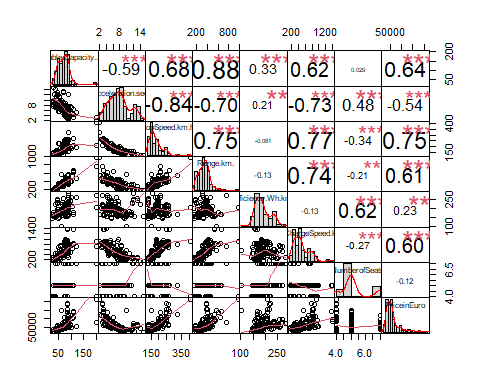


library(psych)  
describe(cars1)

## vars n mean sd median trimmed mad  
## Useable.Capacity..kWh 1 164 66.73 24.40 67.25 65.36 24.39  
## Acceleration.sec. 2 164 7.67 2.97 7.50 7.56 2.97  
## TopSpeed.km.h. 3 164 177.74 42.66 160.00 172.55 37.06  
## Range.km. 4 164 344.09 116.26 340.00 335.95 103.78  
## Efficiency.Wh.km. 5 164 195.35 33.20 189.00 193.09 31.13  
## FastChargeSpeed.km.h. 6 164 486.65 246.97 440.00 460.68 229.80  
## NumberofSeats 7 164 5.24 0.94 5.00 5.17 0.00  
## PriceinEuro 8 164 59644.97 32574.04 50890.00 53645.63 19110.71  
## min max range skew kurtosis se  
## Useable.Capacity..kWh 23.8 200 176.2 1.89 8.83 1.91  
## Acceleration.sec. 2.1 15 12.9 0.34 -0.63 0.23  
## TopSpeed.km.h. 123.0 410 287.0 1.61 4.62 3.33  
## Range.km. 165.0 970 805.0 1.35 4.79 9.08  
## Efficiency.Wh.km. 104.0 281 177.0 0.57 -0.33 2.59  
## FastChargeSpeed.km.h. 120.0 1410 1290.0 1.18 1.58 19.28  
## NumberofSeats 4.0 7 3.0 0.93 -0.12 0.07  
## PriceinEuro 20490.0 215000 194510.0 2.15 5.35 2543.61

Observăm histogramele variabilelor, putem spune că predomină distribuțiile **leptocurtice,** însă aplicând funcția “describe” obtinem valorile pentru skewness si kurtosis care ne arată precis ce tipuri de distribuție avem. Dacă coeficientul de boltire este mai mic decât 3 (<3) avem distribuții platicurtice, în caz contrar leptocurtice ( >3).  
În cazul nostru, distribuții platicurtice avem pentru variabilele: Acceleration(sec), Efficiency.Wh.km., FastCharge, NumberofSeats.  
Indicele de asimetrie pozitiv, ne arată că majoritatea datelor sunt distibuite spre stânga, ceea ce înseamna că predomină valorile mici.

cars2<-cars1[,1:8]  
#cars2  
library(PerformanceAnalytics)  
library(xts)  
library(zoo)  
windows()  
suppressWarnings(chart.Correlation(cars1, histogram=T, pch=19))



Prin pachetul PerformanceAnalytics am generat graficul din figura de mai sus.  
Pe diagonala principală putem observa histogramele pentru fiecare variabilă. Deasupra diagonalei principale regăsim valoarea corelațiilor si totodată nivelul de semnificație al indicatorilor. Fiecare nivel de semnificație are asociat un simbol format din una sau mai multe steluțe. În cazul setului nostru de date, predomină simbolul cu trei stelute, ceea ce înseamnă o valoare a lui P-value de 0.001.  
Sub diagonala principală regăsim legăturile dintre variabilele reprezentate prin grafice de tip plot. Cu ajutorul lor putem analiza legătura dintre fiecare două componente și modul în care sunt distribuite datele de-a lungul bisectoarei care ne dă sensul legăturii.

Spre exemplu, unul dintre ploturile importante regăsite în acest grafic este cel ce face referire la corelația dintre variabilele Useable Capacity: kWh și Range(km), unde putem observa un coeficient de corelație de 0.88, fiind o corelație directă și puternică deoarce tinde spre 1.

## VI. Analiza Componentelor Principale

Pentru a aplica analiza componentelor principale am standardizat datele si am analizat matricea de corelatie pentru a vedea legaturile dintre variabile. Pentru extragerea componentelor principale am utilizat functia princomp pentru datele standardizate si am obtinut urmatorul output:

cars2<-data.frame(cars1[,1:8])  
#View(cars2)  
date\_st<-scale(cars2,scale=T)  
#date\_st  
PCA<-princomp(date\_st,cor = T)  
sdev<-PCA$sdev  
valp<-sdev\*sdev   
procent\_info<-valp\*100/ 8   
procent\_cumulat<-cumsum(procent\_info)  
X<-round(data.frame(sdev,valp,procent\_info,procent\_cumulat),3)  
X

## sdev valp procent\_info procent\_cumulat  
## Comp.1 2.142 4.587 57.333 57.333  
## Comp.2 1.354 1.834 22.921 80.254  
## Comp.3 0.742 0.551 6.886 87.140  
## Comp.4 0.609 0.371 4.636 91.776  
## Comp.5 0.568 0.323 4.038 95.814  
## Comp.6 0.473 0.224 2.803 98.617  
## Comp.7 0.320 0.102 1.276 99.893  
## Comp.8 0.092 0.009 0.107 100.000

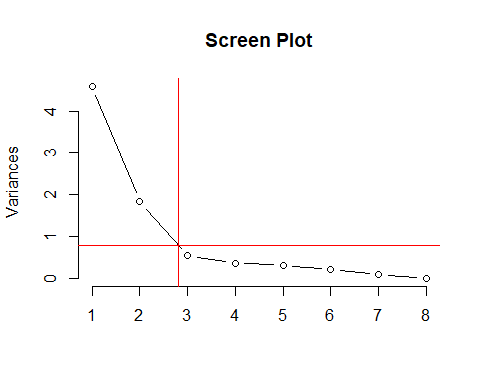
În ceea ce privește prima componentă, abaterea standard a acesteia are cea mai mare valoare dintre cele opt componente identificate, de 2.142. Procentul de informație preluat de aceasta este de 57.333%. De asemenea, în ceea ce privește procentul cumulat de primele două componente este de 80.254%

Pentru a determina numărul de componente principale am urmat 3 criterii de determinare:

* **Criteriul lui Kaiser**, conform căruia se iau în calcul doar numărul de componente principale care au varianta mai mare sau egală cu 1.
* **Criteriul procentului de acoperire** prin care numărul optim de componente este cel care are un procent cumulat mai mare de 80%.
* **Criteriul pantei** prin care se realizează un grafic cu o tăietura după o dreaptă paralelă cu ordonata, astfel încat la dreaptă tăieturii să ramană o porțiune de grafic unde panta este aproape zero.

**Conform celor trei criterii, în analiza trebuie urmărite două componente principale.**

scree\_plot<-prcomp(date\_st)  
plot(scree\_plot, type="l", main="Screen Plot")   
abline(v=2.8, lwd=1.5, col="red")  
abline(h=0.8, lwd=1.5, col="red")



Vom denumi prima componenta **Performanță** – aceasta fiind formata din variabilele Useable Capacity, Acceleration, TopSpeed, Range, FastCharge, PriceinEuro.  
  
Cea de-a doua componenta se va numi **Diverse**– aceasta fiind formata din variabilele Efficiency și NumberOfSeats.

c<-PCA$scores[,1:2]  
  
round(cor(c),3)

## Comp.1 Comp.2  
## Comp.1 1 0  
## Comp.2 0 1

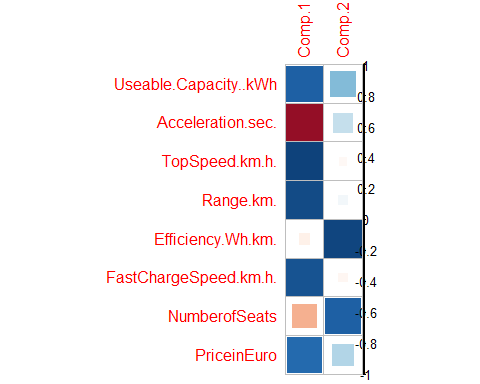
mat\_factor<-cor(date\_st,c)  
round(mat\_factor,3)

## Comp.1 Comp.2  
## Useable.Capacity..kWh 0.819 0.436  
## Acceleration.sec. -0.873 0.235  
## TopSpeed.km.h. 0.930 -0.037  
## Range.km. 0.900 0.058  
## Efficiency.Wh.km. -0.072 0.916  
## FastChargeSpeed.km.h. 0.864 -0.049  
## NumberofSeats -0.354 0.810  
## PriceinEuro 0.775 0.292

library(corrplot)

## corrplot 0.92 loaded

corrplot(mat\_factor,method="square")



library(factoextra)

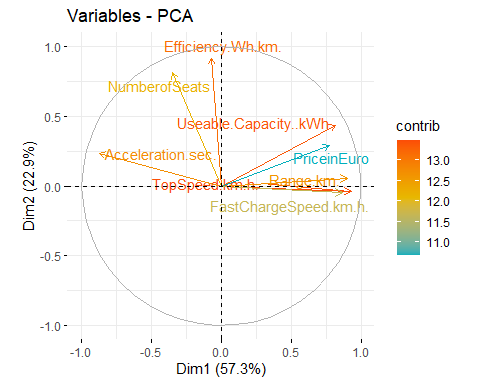
## Loading required package: ggplot2

##   
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(ggplot2)  
  
fviz\_pca\_var(PCA, col.var="contrib",  
 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),  
 repel = TRUE # Avoid text overlapping  
)



Cercul din figura de mai sus ne arată cât de mult au contribuit variabilele analizate la realizarea componentelor principale, unde prima axa este data de prima component principal(Performanță), iar cea de-a doua axă este dată de a doua componentă principală(Diverse). Observăm ca unghiul dintre variabila Efficiency și prima component este foarte ascuțit, ceea ce ne indică un coeficient de corelație cu valoare mare.

## VII. Arbori de clasificare

[Arborele de clasificare de mai jos este determinat prin utilizarea librariei “party” pentru o mai buna organizare a datelor și a codului. Au fost implicate in acesta variabilele “Useable Capacity” și “PriceinEuro”.] [Se observă ca pentru masinile a căror preț este mai mica sau egal cu 35650 euro, capacitatea utilizabiă preponderentă este 40, identificate la 40% dintre masinile analizate si 45 capacitatea utilizabiă , indentificate la 25% dintre mașini. In ceea ce priveste un preț mai mare de 35650 euro, capacitatea utilizabiă preponderentă este de 45, identificată la 20% dintre mașini, in timp ce putin peste 15% variază între 68 și 81.]

set.seed(42)  
n <- nrow(Cars)  
train <- Cars[sample(n, n/4), ]  
  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:xts':  
##   
## first, last

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

train <- sample\_n(Cars, n / 4)  
  
#View(train)  
  
  
library(party)

## Loading required package: grid

## Loading required package: mvtnorm

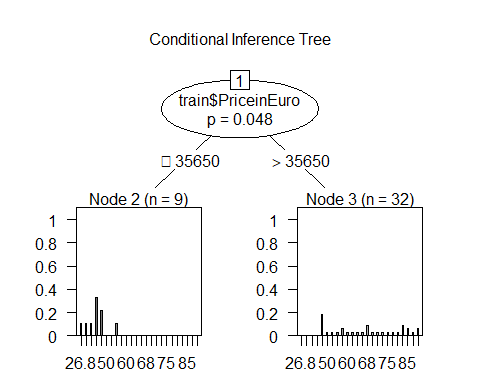
## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: sandwich

ct <- ctree(factor(train$`Useable Capacity: kWh`) ~ train$PriceinEuro, data = train, controls = ctree\_control(minsplit = 3))  
plot(ct, main = "Conditional Inference Tree")



## VIII. Concluzii

În concluzie, în urma analizelor realizate pentru setul de date ales, s-au conturat pricipalele caracteristici ale automobilelor electrice și s-a determinat eficiența mașinilor. În acest sens am determinat 2 componente principale, acestea au fost denumite componenta “Performanță”, fiind formata din variabilele Useable Capacity, Acceleration, TopSpeed, Range, FastCharge, PriceinEuro și componenta “Diverse”, formată din variabilele Efficiency și NumberOfSeats.

Un alt fir analitic urmarit tine de arborii de clasificare, în acest subcapitol s-a observat că pentru mașinile a căror preț este mai mica sau egal cu 35650 euro, capacitatea utilizabiă preponderentă este 40, identificate la 40% dintre masinile analizate si 45 capacitatea utilizabiă , indentificate la 25% dintre mașini. In ceea ce priveste un preț mai mare de 35650 euro, capacitatea utilizabiă preponderentă este de 45, identificată la 20% dintre mașini, in timp ce putin peste 15% variază între 68 și 81.