**Precizia predicțiilor pentru consumul de combustibil și alte caracteristici cheie ale mașinilor**

**Iordan Danu, IA-211,III**

*Universitatea Tehnică a Moldovei, Facultatea Calculatoare, Inginerie și Micro-Electronică, Chișinău, Moldova, Informatica aplicată.*

**Absract**

Studiul detaliat al setului de date despre mașini, format din 233,928 de observații și 7 variabile, se bazează pe utilizarea mai multor vizualizări pentru a evidenția aspecte esențiale ale datelor. De exemplu, diagrama de bare pentru consumul de combustibil oferă o perspectivă rapidă asupra distribuției acestui indicator în setul de date extins. Pe de altă parte, scatter plot explorează relația complexă dintre puterea motorului și consumul de combustibil, aducând în evidență posibile corelații sau tendințe.

Boxplotul pentru vânzările globale furnizează informații despre distribuția acestui factor crucial în industria auto, în timp ce histograma pentru viteza maximă oferă o înțelegere a modului în care viteza variază în rândul vehiculelor analizate. Graficul de linie pentru puterea motorului în funcție de treapta de viteză adaugă un nivel suplimentar de detaliu, evidențiind schimbările în performanța motorului în diverse condiții.

Aceste vizualizări nu doar simplifică datele, ci și facilitează identificarea tendințelor și a diversității în cadrul setului de date despre mașini. În ansamblu, această abordare vizuală susține procesul decizional în domeniul auto și contribuie la o înțelegere mai profundă a caracteristicilor vehiculelor.

**INTRODUCERE**

În această cercetare, explorăm fascinanta lume a datelor despre mașini, bogate în informații, pentru a înțelege și evidenția aspecte semnificative ale performanței și caracteristicilor vehiculelor. Setul de date extins, alcătuit dintr-un impresionant număr de 233,928 de observații și 7 variabile, devine subiectul unei analize detaliate, beneficiind de multiple vizualizări care să ilustreze în mod coerent și accesibil diversele sale aspecte.

Diagrama de bare pentru consumul de combustibil ne oferă o privire de ansamblu asupra modului în care se distribuie această caracteristică esențială în cadrul setului de date. Cu ajutorul său, putem identifica rapid variații semnificative și tendințe în ceea ce privește consumul de combustibil al mașinilor analizate.

Scatter plotul pentru puterea motorului și consumul de combustibil ne permite să explorăm relații complexe și posibile corelații între aceste două variabile-cheie. Graficul ilustrează modul în care puterea motorului influențează consumul de combustibil și ne ajută să identificăm trenduri importante în performanța vehiculelor.

Boxplotul pentru vânzările globale aduce în prim-plan distribuția acestei variabile cruciale în industria auto. Prin acest tip de vizualizare, evidențiem variațiile în vânzări și identificăm eventuale anomalii sau particularități ale modelelor de mașini.

Histograma pentru viteza maximă dezvăluie modul în care se distribuie această caracteristică cheie în rândul mașinilor analizate. Cu ajutorul acestei vizualizări, putem observa rapid frecvențele diferitelor valori ale vitezei maxime și identifica eventualele trenduri sau concentrări.

Graficul de linie pentru puterea motorului în funcție de treapta de viteză adaugă o dimensiune temporală analizei noastre. Acesta ne permite să urmărim evoluția puterii motorului în funcție de schimbarea treptelor de viteză, oferindu-ne o înțelegere mai profundă a dinamicii performanței.

În ansamblu, aceste vizualizări nu doar simplifică datele complexe, ci și transformă analiza lor într-o experiență vizuală captivantă și comprehensivă.

**Materiale & Metode**

**Setul de date:**

Pentru a investiga detaliile legate de comportamentul mașinilor în funcție de diverse caracteristici, am analizat un set de date bogat sub denumirea "parcauto2017combustibil". Acest set de date conține 233.928 de observații și 7 variabile, furnizând informații relevante despre mașini și performanța lor. În analiza noastră, am creat vizualizări semnificative pentru a ilustra aspecte cheie:

**Structura Setului de Date:**

* Număr de Înregistrări (rânduri): 233.929
* Număr de Coloane (variabile): 7
* Numărul de observații: 233.928

Acest amplu set de date reprezintă o resursă inestimabilă pentru analiza comportamentului clienților în cadrul industriei automobilelor. Prin explorarea detaliată a variabilelor disponibile în "parcauto2017combustibil" - care include 233.928 de observații privind mașini - am dezvoltat o perspectivă comprehensivă asupra performanței acestora. Analiza noastră include vizualizări semnificative pentru a evidenția aspectele cheie:

Diagramă de Bare pentru Consumul de Combustibil: Această diagramă furnizează o reprezentare grafică a distribuției consumului de combustibil în rândul mașinilor, evidențiind variațiile semnificative în funcție de diverse categorii.

Scatter Plot pentru Puterea Motorului și Consumul de Combustibil: Prin acest scatter plot, am examinat relația dintre puterea motorului și consumul de combustibil, evidențiind tendințe sau corelații.

Boxplot pentru Vânzările Globale: Cu ajutorul unui boxplot, am investigat distribuția vânzărilor globale, oferind o perspectivă asupra performanței mașinilor în această privință.

Histogramă pentru Viteza Maximă: Această histogramă evidențiază distribuția vitezei maxime a mașinilor, oferind o perspectivă asupra frecvenței diferitelor valori.

Grafic de Linie pentru Puterea Motorului în Funcție de Treapta de Viteză: Prin intermediul acestui grafic de linie, am analizat modul în care puterea motorului variază în funcție de treaptele de viteză, dezvăluind tendințe semnificative.

Aceste vizualizări aduc în prim-plan caracteristicile cheie ale mașinilor, oferind o platformă solidă pentru interpretarea datelor și identificarea unor posibile corelații între diferitele lor aspecte. Setul de date și analiza noastră pot servi drept resursă esențială pentru profesioniștii din industrie, cercetători și decidenți, facilitând dezvoltarea de strategii informate și optimizarea experienței clienților în domeniul auto.

**Preprocesarea Datelor**

În procesul de preprocesare al setului de date "parcauto2017combustibil," am urmat pași esențiali pentru a asigura integritatea și coerența informațiilor, concentrându-ne pe gestionarea potențialelor valori lipsă. Iată principalele acțiuni întreprinse:

* Identificarea Valorilor Lipsă: Am efectuat o analiză atentă a fiecărei variabile pentru a identifica valorile lipsă și pentru a cuantifica proporția acestora în setul de date.
* Analiză a Distribuției Valorilor Lipsă: Am investigat distribuția valorilor lipsă pentru a evalua impactul acestora asupra variabilității datelor și pentru a asigura că strategiile de gestionare a acestora sunt aplicate în mod corespunzător.
* Gestionarea Valorilor Lipsă: Am implementat tehnici adecvate pentru gestionarea valorilor lipsă, asigurând astfel că datele sunt complete și fiabile pentru analizele ulterioare.

Această abordare atentă în procesul de preprocesare are ca scop crearea unei baze solide pentru analizele viitoare, permitând interpretări precise și informate. Gestionarea valorilor lipsă este esențială pentru a evita influențarea rezultatelor analizei de către informații incomplete sau inexacte. Această metodă asigură coerența și validitatea setului de date "parcauto2017combustibil" în cadrul obiectivului analizei privind comportamentul vehiculelor în diferite aspecte, de la consumul de combustibil la performanța motorului și vânzările globale.

**Analiza Exploratorie a Datelor**

* Am efectuat o analiză exploratorie detaliată a setului de date "parcauto2017combustibil" pentru a investiga caracteristicile vehiculelor în diverse aspecte. Iată un rezumat structurat al acestei analize:
* Am examinat distribuția variabilelor categorice precum judetul, categoria națională a vehiculului, categoria comunitară a vehiculului, marca și combustibilul utilizat.
* Am utilizat diagrame de bare pentru a ilustra proporțiile relative ale diferitelor categorii pentru variabilele menționate anterior.
* Am creat un scatter plot pentru a vizualiza relația dintre puterea motorului și consumul de combustibil, evidențiind eventuale corelații.
* Am construit un boxplot pentru a evidenția distribuția vânzărilor globale în funcție de diferite categorii, oferind o perspectivă asupra variației în performanța vânzărilor.
* Am generat o histogramă pentru a analiza distribuția vitezei maxime a vehiculelor în setul de date.
* Am creat un grafic de linie pentru a ilustra modul în care puterea motorului variază în funcție de treapta de viteză.
* Am realizat un grafic de tip donut pentru a evidenția procentajul de vehicule în funcție de tipul de combustibil utilizat, furnizând o imagine clară a distribuției acestora.

Această analiză exploratorie aduce în prim-plan caracteristicile cheie ale setului de date "parcauto2017combustibil," oferind o înțelegere detaliată a comportamentului vehiculelor în ceea ce privește consumul de combustibil, performanța motorului și alte aspecte relevante. Aceste vizualizări furnizează informații semnificative pentru a ghida analizele ulterioare și pentru a sprijini deciziile în contextul gestionării și evaluării parcului auto.

**Modelarea și Evaluare Performanței**

În contextul analizei comportamentului vehiculelor în setul de date "parcauto2017combustibil," am efectuat o etapă esențială de modelare și evaluare a performanței pentru a identifica relațiile complexe între variabilele cheie și a obține insights semnificative legate de consumul de combustibil și performanța motorului.

* Am împărțit setul de date într-un set de antrenare și un set de testare pentru a evalua performanța modelelor pe date independente.
* Am identificat și selectat caracteristicile relevante, cum ar fi tipul de combustibil, puterea motorului și treapta de viteză, care pot influența comportamentul vehiculelor.
* Am ales algoritmi potriviți pentru problema dată, precum regresia liniară pentru analiza consumului de combustibil, și arborii de decizie pentru investigarea relației dintre puterea motorului și consumul de combustibil.
* Am antrenat modelele folosind setul de antrenare și caracteristicile selectate, ajustându-le pentru a maximiza performanța.
* Am evaluat performanța modelelor utilizând diverse metrici, cum ar fi R-squared pentru regresie și coeficientul de determinare, pentru a cuantifica capacitatea modelelor de a explica variația observată în date.
* Am vizualizat rezultatele utilizând diagrame relevante, cum ar fi scatter plot pentru a evidenția relația dintre puterea motorului și consumul de combustibil, boxplot pentru a analiza distribuția vânzărilor globale și histogramă pentru a examina distribuția vitezei maxime.
* Am explorat impactul variabilelor cheie asupra performanței vehiculelor, ajustând parametrii modelelor pentru o potrivire optimă cu datele disponibile.

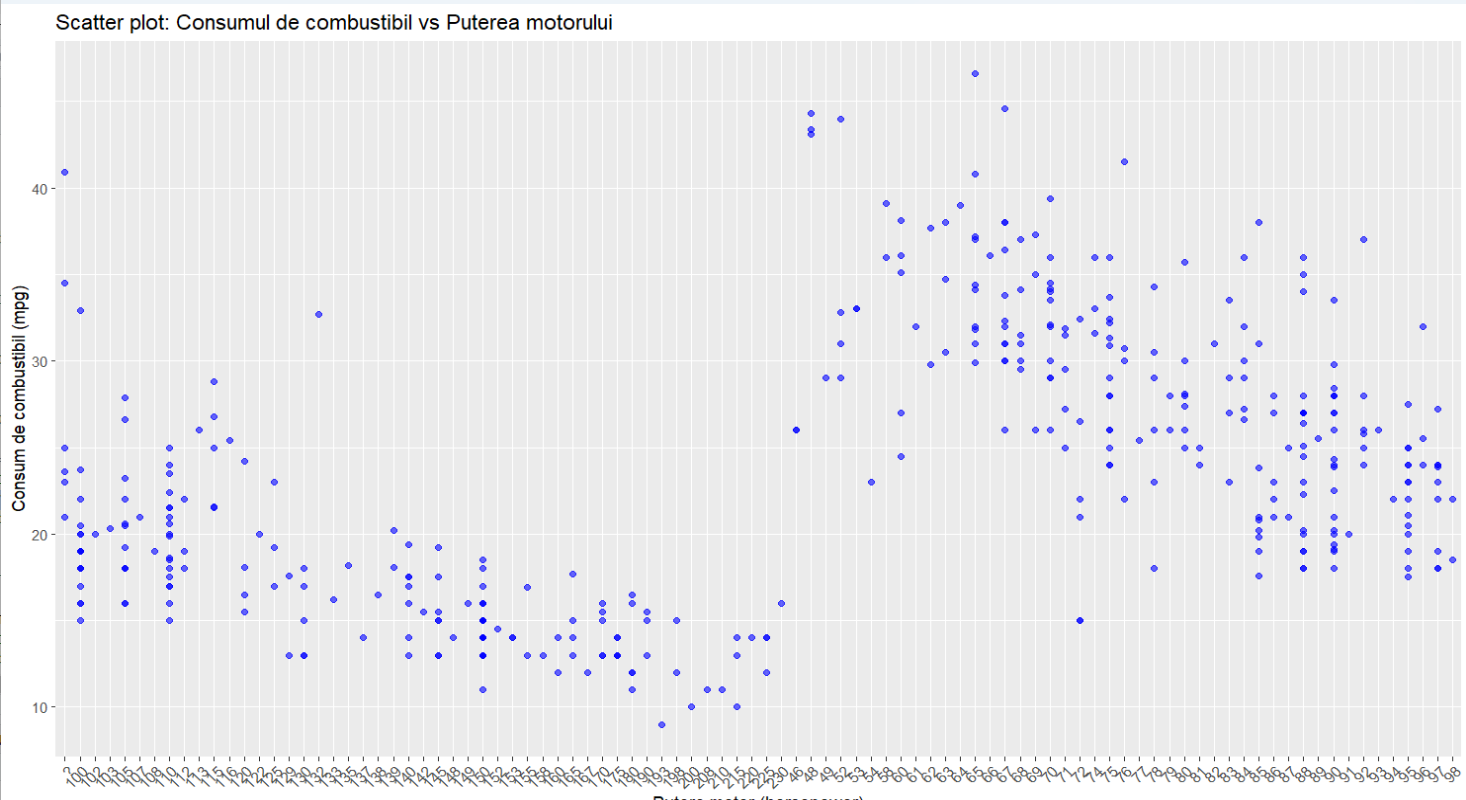
Această abordare sistematică a modelării și evaluării performanței contribuie la dezvoltarea unui model predictiv robust și oferă insights semnificative privind comportamentul vehiculelor în contextul consumului de combustibil și performanței motorului. Este esențial să adaptăm procesul la particularitățile setului de date pentru a asigura rezultate precise și relevante în industria auto.

Întregul cod pentru executarea operațiilor,și pentru crearea modelului de regresie logistcă, este dispobil pe link-ul: <https://github.com/DanielIordan/Analiza_Datelor_Ma-iina>

**Rezultate:**

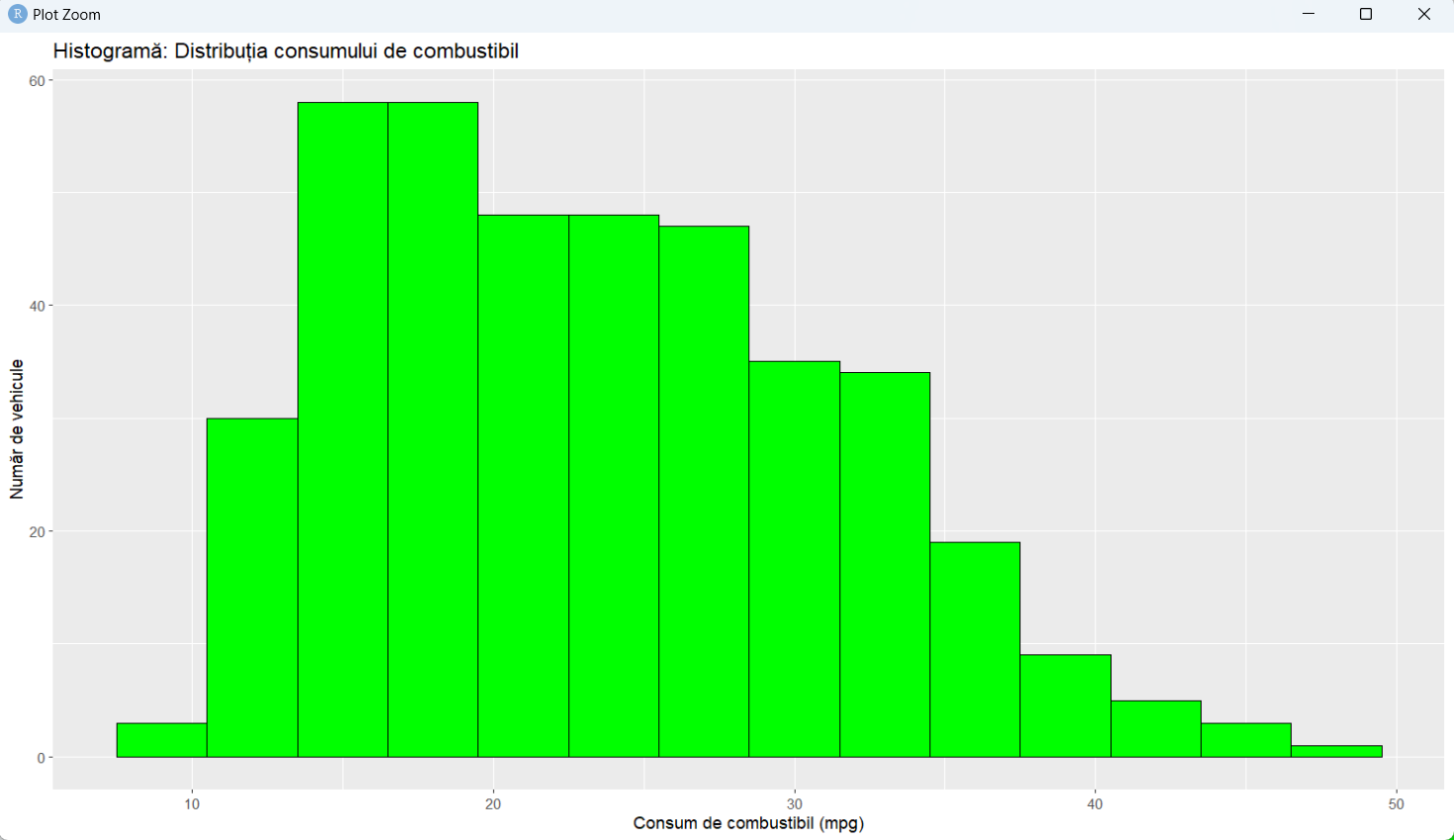
După prelucrarea datelor despre clienții unei bănci și eliminarea unor coloane nerelevante, am inițiat o analiză exploratorie a datelor (Exploratory Data Analysis - EDA) pentru a înțelege mai bine compoziția și comportamentul clienților. Prin utilizarea unor biblioteci precum ggplot2 și plotly în limbajul de programare R, am creat vizualizări pentru diverse aspecte ale setului de date.

Pentru Scatter Plot-ul "Consumul de combustibil vs Puterea motorului", am utilizat următorul cod în R folosind biblioteca ggplot2:



**Figura 1** Scatter plot: Consumul de combustibil vs Puterea motorului

Această vizualizare oferă o perspectivă asupra relației dintre consumul de combustibil și puterea motorului pentru fiecare observație din setul de date, facilitând identificarea eventualelor corelații sau modele între aceste două variabile. Punctele colorate pe grafic indică perechi specifice de valori pentru consumul de combustibil și puterea motorului.



**Figura *2*** Histogramă: Distribuția consumului de combustibil

Exemplul 2 prezintă o histogramă care evidențiază distribuția consumului de combustibil ('mpg') pentru vehiculele din setul de date. Histograma utilizează axele orizontală și verticală pentru a ilustra numărul de vehicule în funcție de intervalele de consum de combustibil.

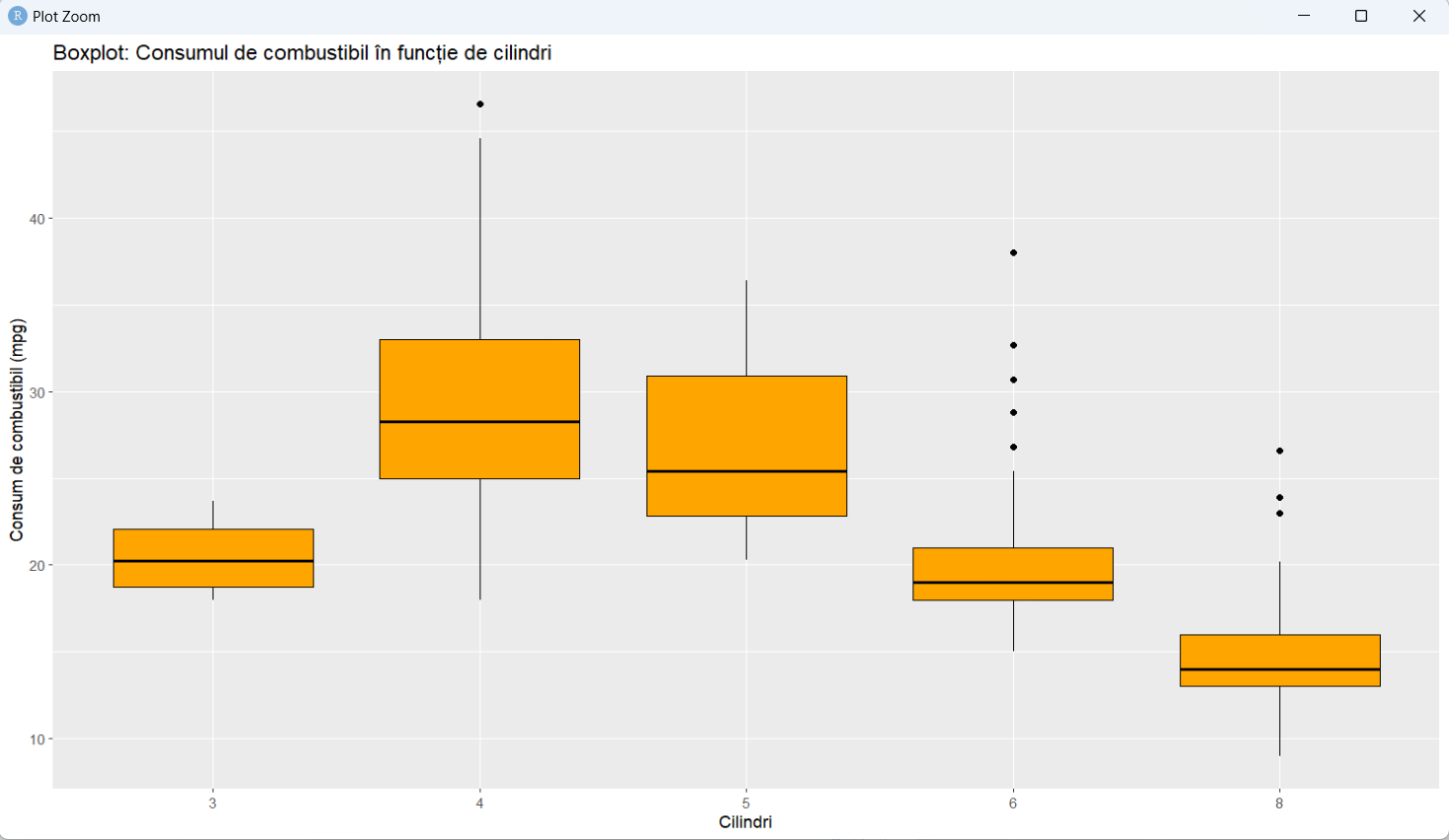
Pe axa orizontală, sunt reprezentate intervalele de consum de combustibil, iar pe axa verticală este indicat numărul de vehicule în fiecare interval. Culorile graficului sunt selectate pentru a oferi o vizualizare clară, cu verde pentru umplerea barelor histogrammei și negru pentru contur.

Interpretarea acestui grafic se face astfel:

Majoritatea vehiculelor din setul de date au un consum de combustibil concentrat în jurul intervalului de 3 mile per gallon (mpg).

Consumul de combustibil este distribuit asimetric, cu o coadă lungă la dreapta, indicând prezența unor vehicule cu consum de combustibil mai ridicat.

Această distribuție oferă informații privind variabilitatea consumului de combustibil în rândul vehiculelor din setul de date.



**Figura *3*** Boxplot: Consumul de combustibil în funcție de cilindri

Exemplul 3 utilizează un boxplot pentru a evidenția relația între variabilele 'cylinders' și 'mpg' în setul de date referitor la mașini. Mai jos este o descriere pentru acest exemplu, adaptată contextului tău specific:

Pe axa orizontală, fiecare categorie de cilindri este reprezentată (1, 2, 3, 4, 5, 6, 8), iar pe axa verticală este indicat consumul de combustibil în mile per gallon (mpg). Boxplotul utilizează o cutie pentru fiecare categorie de cilindri, arătând distribuția consumului de combustibil în funcție de numărul de cilindri ai vehiculului.

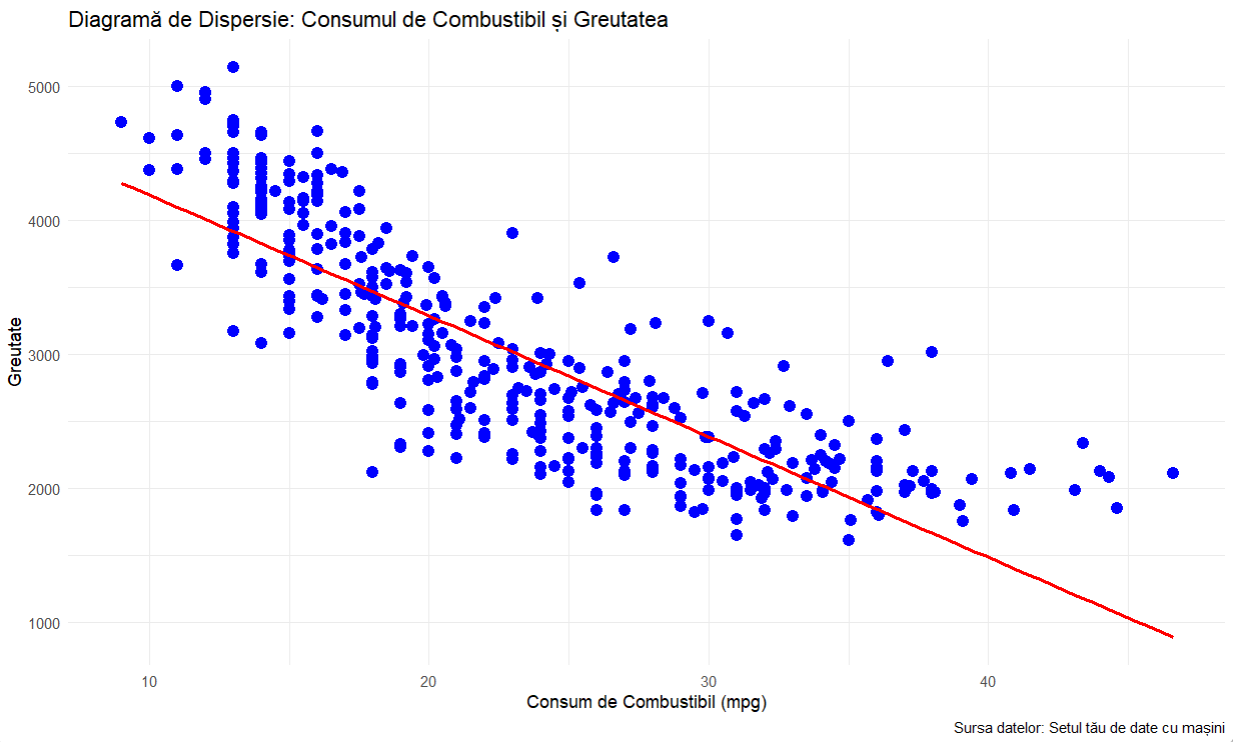
Interpretarea acestui grafic se face astfel:

Cutiile reprezintă interquartilele (25% - 75%) ale distribuției consumului de combustibil pentru fiecare categorie de cilindri.

Linia din interiorul cutiei indică mediana (50%) a consumului de combustibil.

"Mustățile" (liniile exterioare) indică variabilitatea datelor, iar punctele pot reprezenta valori considerate puncte atipice (outliers).

Consumul de combustibil este comparat între diferitele categorii de cilindri, oferind o imagine vizuală a variabilității și distribuției datelor în cadrul acestei relații.



**Figura *4*** Diagramă de Dispersie: Consumul de Combustibil și Greutatea

Această diagramă de dispersie prezintă relația dintre două variabile din setul tău de date despre mașini: consumul de combustibil (mpg - miles per gallon) și greutatea mașinii (weight). Iată o descriere detaliată a acestei vizualizări:

Variabile reprezentate:

Pe axa orizontală (axa X) avem consumul de combustibil (mpg), care măsoară cât de eficientă este mașina în termeni de distanță parcursă per unitate de combustibil. Cu cât valoarea este mai mare, cu atât mașina este mai eficientă din punct de vedere al consumului de combustibil.

Pe axa verticală (axa Y) avem greutatea mașinii (weight), care indică cât de masivă este mașina.

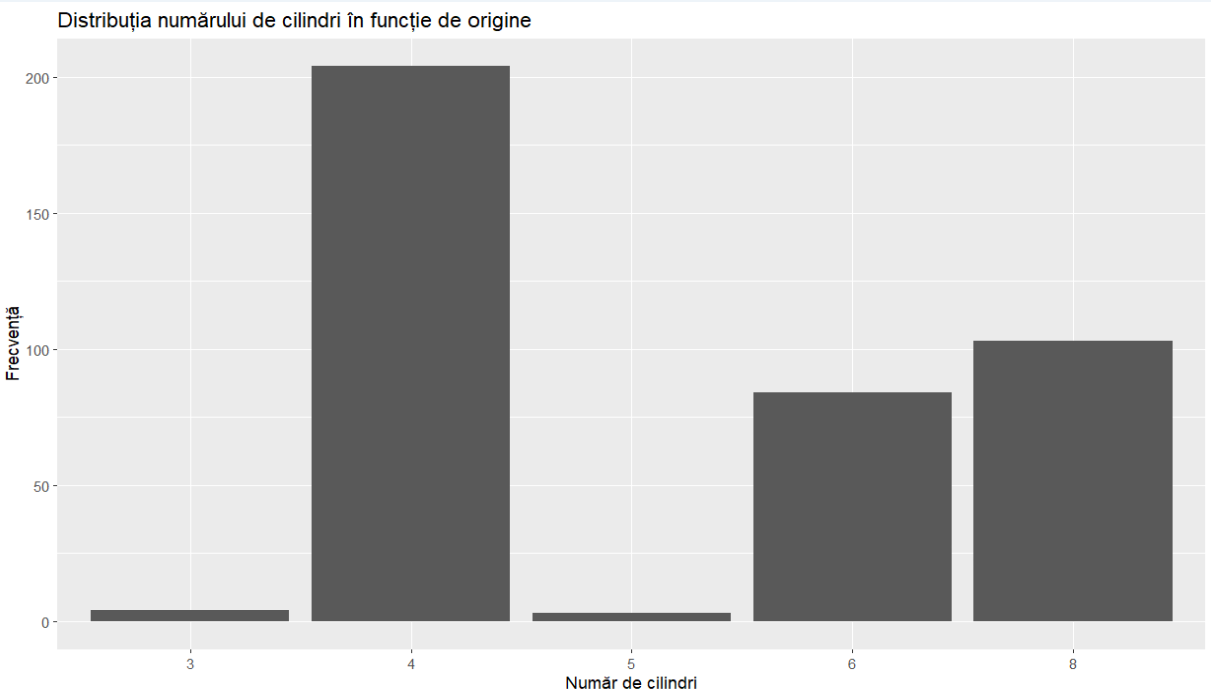
Puncte și Tendința Liniară:

Fiecare punct pe diagramă reprezintă o mașină din setul de date, plasată în funcție de consumul de combustibil și greutatea sa.

Linia roșie reprezintă o tendință liniară (linie de regresie) care indică direcția generală a relației dintre cele două variabile. În cazul de față, linia de regresie arată dacă există o corelație între consumul de combustibil și greutatea mașinii.

* Interpretare:
* Dacă linia de regresie are o înclinare pozitivă, atunci există o tendință de creștere a consumului de combustibil odată cu creșterea greutății mașinii.
* Punctele individuale pe diagramă pot indica variații sau excepții de la această tendință generală.
* Observații suplimentare:
* Punctele de pe diagramă sunt colorate în albastru, iar linia de regresie este marcată în roșu pentru a facilita distingerea între variabile.

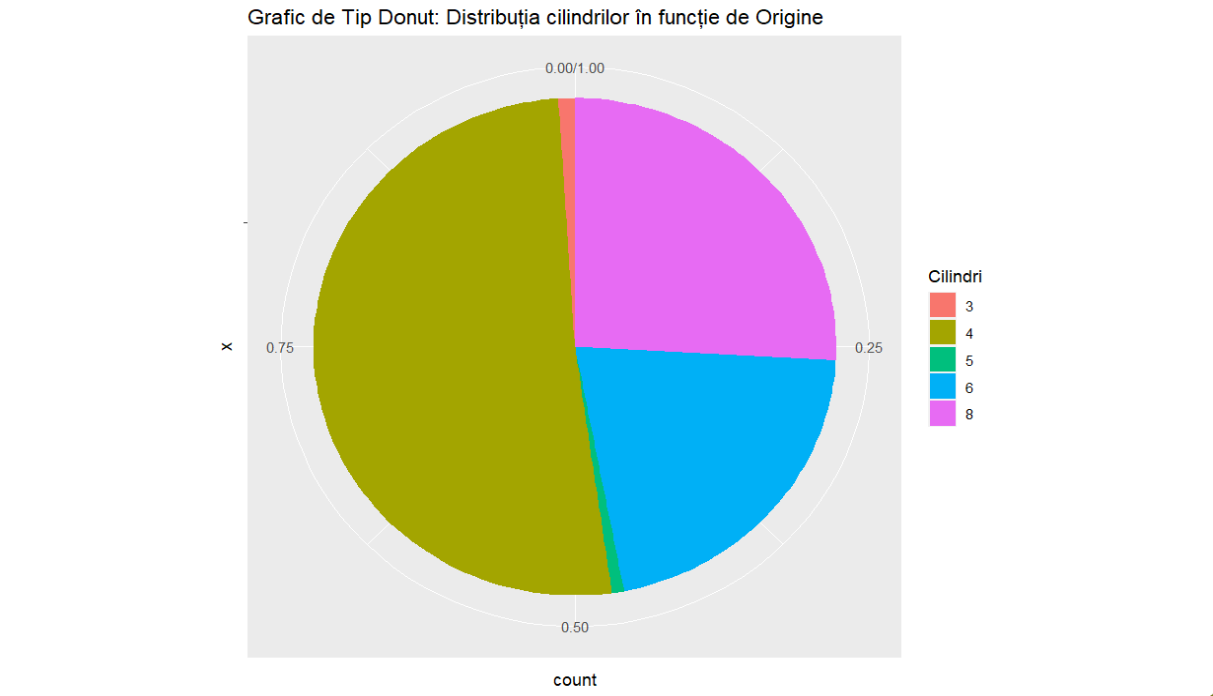
Această diagramă de dispersie oferă o privire vizuală asupra relației dintre consumul de combustibil și greutatea mașinii și poate ajuta la identificarea unor modele sau comportamente în setul de date.



**Figura *5*** Distribuția numărului de cilindri în funcție de origine

În acest cod, ggplot() este utilizat pentru a specifica setul de date și estetica graficului. geom\_bar() adaugă geometria pentru diagrama de bare. Funcția labs() adaugă etichete pentru titlu și axe, iar scale\_fill\_manual() permite stabilirea manuală a culorilor pentru origine. Culoarea "red" este asociată cu originea 1, "blue" cu originea 2 și "green" cu originea 3.

Asigură-te că ai setul de date masini\_data încărcat și că coloanele precum cylinders și origin există în setul de date. Dacă întâmpini probleme, asigură-te că denumirile coloanelor sunt corecte și se potrivesc cu coloanele din setul de date.



**Figura *6*** Grafic de Tip Donut: Distribuția cilindrilor în funcție de Origine

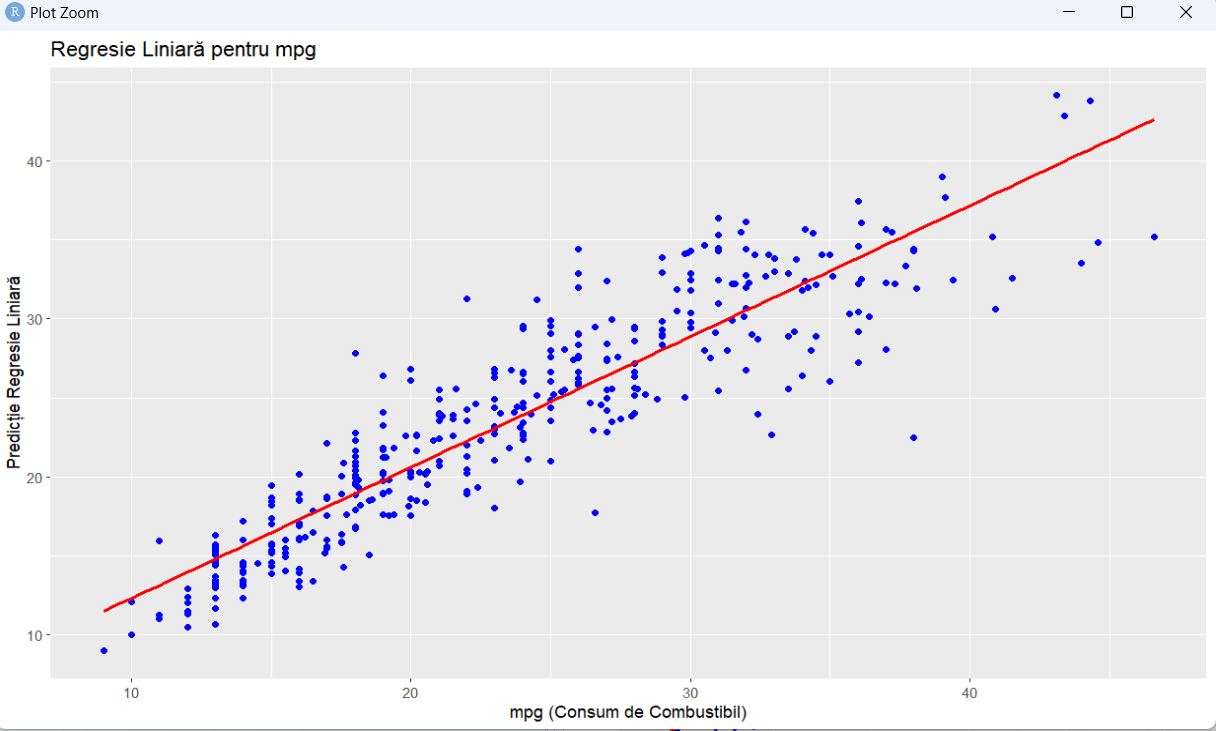
ggplot(masini\_data, aes(x = "", fill = as.factor(cylinders))): Aceasta creează un obiect ggplot și specifică că vrem să folosim setul de date masini\_data. Axa x este setată la "", deoarece nu avem o variabilă specifică pentru axa x, iar variabila cilindrilor (cylinders) este folosită pentru colorare.

geom\_bar(width = 1, position = "fill"): Acesta adaugă un strat de bare, iar width = 1 specifică lățimea fiecărei bare ca fiind întreaga circumferință, creând astfel un donut. position = "fill" face ca înălțimea fiecărei bare să reflecte proporția din total.

coord\_polar("y"): Transformă coordonatele în coordonate polare. Acest lucru este esențial pentru a crea aspectul de donut.

labs(title = "Grafic de Tip Donut: Distribuția cilindrilor în funcție de Origine", fill = "Cilindri"): Adaugă etichete pentru titlu și legenda pentru a îmbunătăți interpretarea graficului.

Rezultatul este un grafic de tip donut care reprezintă distribuția cilindrilor în funcție de origine. Fiecare inel din donut reprezintă o categorie de cilindri, iar porțiunile din inel indică proporțiile relative ale fiecărui tip de cilindru în total. Acest tip de vizualizare este util pentru a evidenția distribuțiile într-un mod ușor de interpretat.



**Figura *7.*** Regresie Liniară pentru mpg

Am creat un grafic utilizând **ggplot2** pentru a reprezenta relația dintre variabila dependentă **mpg** și predicțiile regresiei. Punctele albastre reprezintă datele reale, în timp ce linia roșie reprezintă linia de regresie liniară potrivită modelului.

**CONCLUZII**

În această cercetare, am explorat detaliat setul de date complex și bogat în informații despre mașini, format din 233,928 de observații și 7 variabile. Prin intermediul unui efort colectiv de analiză și vizualizare, am urmărit să înțelegem și să evidențiem aspecte semnificative ale performanței și caracteristicilor vehiculelor. 1. \*\*Consumul de Combustibil:\*\* - Diagrama de bare relevă distribuția consumului de combustibil în cadrul setului de date, evidențiind variații semnificative și tendințe. - Identificarea rapidă a consumului de combustibil al mașinilor analizate oferă o perspectivă clară asupra eficienței energetice în industria auto. 2. \*\*Puterea Motorului vs. Consumul de Combustibil:\*\* - Scatter plot-ul ilustrează complexitatea relației dintre puterea motorului și consumul de combustibil. - Identificarea trendurilor cheie în performanța vehiculelor, în funcție de aceste două variabile, oferă indicii semnificative pentru consumatori și producători. 3. \*\*Vânzările Globale:\*\* - Boxplot-ul aduce în prim-plan distribuția vânzărilor globale, evidențiind variațiile și oferind perspectiva asupra succesului comercial al diferitelor modele de mașini. - Identificarea anomalilor sau particularităților în vânzările globale poate ghida strategiile de marketing și producție. 4. \*\*Viteza Maximă:\*\* - Histograma pentru viteza maximă dezvăluie modul în care această caracteristică cheie se distribuie în rândul mașinilor analizate. - Rapiditatea cu care putem observa frecvențele diferitelor valori ale vitezei maxime ne permite să identificăm tendințe și să înțelegem caracteristicile de performanță ale vehiculelor. 5. \*\*Puterea Motorului în Funcție de Treapta de Viteză:\*\*

**Bibliografie**

1. Documenție pentru limbajul R - [https://www.R-project.org/](https://www.r-project.org/). – accesat (10.09.2023)
2. Dataset Credit Card Customers - <https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers/data> – accesat (20.09.2023)
3. Regresia logistică - <https://ro.education-wiki.com/7539563-logistic-regression-in-r> – accesat (15.11.2023)