**REFERAT**

**Tehnici de vizualizarea a datelor pentru analiza Big Data**

**MASTER, BAZE DE DATE ȘI TEHNOLOGII SOFTWARE**

**Anul I, grupa 405, Caragea Anda-Maria**

CUPRINS

[**ABSTRACT** 3](#_Toc135419078)

[**INTRODUCERE** 3](#_Toc135419079)

[**TEHNICI DE VIZUALIZARE A DATELOR** 4](#_Toc135419080)

[**1.** **LINE CHART** 4](#_Toc135419081)

[**2.** **MAP** 6](#_Toc135419082)

[**3.** **BAR CHART** 7](#_Toc135419083)

[**4.** **DONUT CHART** 8](#_Toc135419084)

[**5.** **BUTTERFLY PLOT** 9](#_Toc135419085)

[**6.** **SCATTER PLOT** 10](#_Toc135419086)

[**DASHBOARD** 11](#_Toc135419087)

[**CONCLUZII** 12](#_Toc135419088)

[**BIBLIOGRAFIE** 13](#_Toc135419089)

# **ABSTRACT**

În acest referat, s-au explorat tehnici de vizualizare a datelor și importanța vizualizării datelor în contextul analizei Big Data. Tehnicile de vizualizare a datelor joacă un rol cheie în abordarea acestei provocări, oferind un mijloc puternic de înțelegere și analiza a Big Data.

Lucrarea se concentrează pe utilizarea platformei Tableau pentru aplicarea tehnicilor și vizualizarea datelor. Tableau este un instrument interactiv pentru analiza și prezentarea datelor. Diferitele tipuri de grafice și diagrame conduc la obținerea unei înțelegeri cuprinzătoare a datelor, identificând modele, tendințe și relații semnificative.

Exemplele prezentate în lucrare includ utilizarea diagramei cu linii, diagramei pe coloane, graficului cu bare și a altor tipuri de diagrame relevante în vizualizarea datelor. Aceste tehnici permit o prezentare eficientă a evoluției, a relațiilor dintre variabile, a distribuției, facilitând identificarea tendințelor semnificative, a modelelor.

Utilizarea diagramei cu linii pentru analiza tendințelor și relațiilor dintre variabile, a hărții pentru generarea distribuției sau proporției datelor în fiecare regiune, a diagramei cu bare pentru comparația valorilor discrete și a diagramei gogoașă pentru distribuția vânzărilor în diferite regiuni demonstrează importanța tehnicilor de vizualizare a datelor pentru analiza Big Data. În plus, cu ajutorul butterfly plot și scatter plot, instrumente utile pentru vizualizarea și analizarea datelor în cadrul Big Data, am obținut o înțelegere mai profundă. Utilizând butterfly plot, am comparat veniturile generate de genul feminin și genul masculin în diferite categorii de produse. Acest lucru oferă informații cu privire la preferințele și modelele de consum ale celor două grupuri. Scatter plot, pe de altă parte, a ajutat la examinarea relației între două variabile continue, prețul și cantitatea de produse vândute. Acest grafic a ajutat la identificarea corelației între aceste două variabile.

Lucrarea evidențiază beneficiile vizualizării datelor în procesul de luare a deciziilor și în dezvoltarea de strategii bazate pe date. Prin transpunerea datelor în forme vizuale, informațiile complexe devin mai ușor de interpretat, de înțeles.

# **INTRODUCERE**

În lumea actuală, bazată pe date, capacitatea de a da sens unor cantități mari de informații este crucială pentru companii, organizații și cercetători. Odată cu creșterea Big Data, provocarea constă nu numai în colectarea și stocarea seturilor de date masive, ci și în extragerea informațiilor valoroase din acestea. Tehnicile de vizualizare a datelor reprezintă un factor cheie în abordarea acestei provocări. Reprezentând date complexe în formate vizuale, aceste tehnici oferă un mijloc puternic de înțelegere și analiză Big Data.

Domeniul vizualizării datelor a evoluat semnificativ în ultimii ani, determinat de progresele tehnologice și de cererea tot mai mare de modalități eficiente de a comunica informații complexe. Vizualizările ajută la descoperirea tiparelor, tendințelor și relațiilor care altfel ar putea rămâne ascunse în marea vastă de date. Ele facilitează o înțelegere intuitivă, permițând factorilor de decizie să facă alegeri informate și să obțină informații utile. Prin intermediul acestor tehnici, datele sunt transpuse în forme vizuale ce facilitează înțelegerea rapidă și cuprinzătoare a informațiilor complexe, permițându-ne să identificăm modele ascunse, anomalii sau tendințe semnificative. De asemenea, ele ne ajută să comunicăm și să prezentăm rezultatele analizei într-un mod accesibil și captivant, aducând beneficii atât specialiștilor în domeniul Big Data, cât și decidenților și publicului larg.

În cadrul acestui referat, am explorat diverse tehnici de vizualizare a datelor adaptate pentru analiza Big Data, concentrându-mă pe vizualizări specifice prezentate cu ajutorul platformei Tableau. Prin intermediul diferitelor tipuri de grafice și diagrame care permit prezentarea în mod eficient a distribuției, evoluției și relațiilor dintre variabile, se obține o înțelegere cuprinzătoare a datelor, permițând identificarea modelelor, tendințelor și perspectivelor semnificative. În mod specific, vor fi examinate utilizarea diagramei cu linii, diagramei pe coloane, graficul cu bare și a altor tipuri de diagrame relevante în vizualizarea creată în Tableau.

Prin examinarea acestor tehnici și prin analiza exemplelor practice, se evidențiază beneficiile și importanța vizualizării datelor în analiza Big Data. Este prezentată influența pozitivă pe care o au aceste tehnici în procesul de luare a deciziilor, în identificarea de oportunități și în dezvoltarea de strategii bazate pe date. Astfel, se va demonstra că vizualizarea datelor nu este doar o modalitate estetică de reprezentare, ci o unealtă esențială pentru înțelegerea, interpretarea și valorificarea Big Data.

Setul de date pe care l-am ales conține informații despre comenzile plasate de către clienți. Fiecare rând din setul de date reprezintă o singură comandă, iar coloanele oferă detalii despre comandă, cum ar fi identificatorul ei unic, anume ID-ul comenzii, data la care a fost plasată comanda, starea curentă, ID-ul articolului, cantitatea de articole comandate, valoarea totală a comenzii și alte date reprezentative. Analizând acest set de date, se pot obține foarte multe informații, precum: comenzile efectuate, stările și articolele specifice comandate. În plus, se pot analiza prețurile, cantitățile, reducerile aplicate, pentru a înțelege aspectele financiare ale comenzilor.

Pentru vizualizarea datelor am folosit Tableau, un instrument popular de vizualizare a datelor folosit pentru analiza și prezentarea datelor, inclusiv a seturilor de date masive. Tableau oferă o interfață ușor de utilizat și interactivă. Este conceput pentru a gestiona volume mari de date și se poate conecta la diverse surse de date. Oferă un set cuprinzător de instrumente și caracteristici de analiză care permit efectuarea calculelor avansate, analize statistice, prognoză și modelare predictivă. Așadar, prin intermediul Tableau, am explorat și obținut într-un mod eficient informații din setul de date ales.

# **TEHNICI DE VIZUALIZARE A DATELOR**

## **LINE CHART**

Line charts sau diagramele cu linii sunt un tip de vizualizare a datelor care afișează puncte de date conectate prin linii drepte. Ele sunt utilizate în mod obișnuit pentru a arăta relația și tendințele dintre două variabile, de obicei reprezentate pe axa X și pe axa Y. În contextul analizei Big Data, diagramele cu linii joacă un rol semnificativ în înțelegerea și interpretarea seturilor mari de date. Pot dezvălui relații și corelații între diferite variabile, ceea ce ajută la identificarea relațiilor cauză-efect și la luarea deciziilor bazate pe date. Ajută și la prognozarea tendințelor viitoare și la realizarea de predicții bazate pe date istorice. Acest tip de diagramă simplifică prezentarea datelor vaste și complexe prin condensarea lor într-un format vizual, care este mai ușor de înțeles și de interpretat. Acest tip de grafic îmbunătățește comunicarea și facilitează luarea deciziilor bazate pe date. Se pot prezenta constatări complexe într-o manieră atrăgătoare și intuitivă din punct de vedere vizual, facilitând înțelegerea și interpretarea datelor. Sprijină luarea deciziilor și oferă o bază pentru dezvoltarea strategiilor bazate pe date.

Graficul de mai jos reprezintă venitul lunar din octombrie 2020 până în septembrie 2021. Fiecare punct reprezintă valoarea veniturilor în milioane de dolari pentru o anumită lună.

A picture containing line, diagram

Description automatically generated

Pe baza acestui grafic, se pot identifica cu ușurință lunile cu venituri mari, putându-se aloca mai multe resurse în acele perioade pentru a maximiza profitabilitatea. Se poate observa o creștere bruscă a veniturilor în decembrie 2020, dar și în aprilie 2021, indicând o perioadă semnificativă de vânzări. Această perspectivă poate influența deciziile legate de gestionarea stocurilor, personalul și marketingul în lunile cu venituri de vârf. Creșterea vizibilă a veniturilor în decembrie 2020, indică o oportunitate potențială de valorificare a vânzărilor de sărbători sau a campaniilor promoționale în cursul lunii respective.

## **MAP**

Vizualizarea hărților este utilizată pentru a analiza și afișa datele geografic și pentru a le prezenta sub formă de hărți. Acest tip de exprimare a datelor este mai clară și mai intuitivă. Putem vedea vizual distribuția sau proporția datelor în fiecare regiune. Este convenabil pentru toată lumea să extragă informații mai profunde și să ia decizii mai bune. Există multe tipuri de vizualizare a hărților, cum ar fi hărți administrative, hărți termice, hărți statistice, hărți de traiectorie, etc.

O hartă regională este o tehnică de vizualizare utilizată pentru a reprezenta date bazate pe regiuni geografice. Afișează date folosind diferite culori sau modele de umbrire pentru a reprezenta variațiile unei anumite variabile în diferite regiuni. Hărțile regionale sunt utilizate în mod obișnuit în analiza Big Data pentru a înțelege și a analiza datele într-un context geografic. Permit analiza geospațială, unde datele sunt analizate pe baza locației și a proximității de alte caracteristici. Prin vizualizarea datelor pe o hartă regională, analiștii pot identifica modele spațiale, clustere sau anomalii care ar putea să nu fie evidente în formatele tradiționale tabelare sau numerice. Evidențiază disparitățile regionale prin reprezentarea vizuală a variațiilor datelor în diferite regiuni. Acest lucru este deosebit de util în analiza Big Data, deoarece le permite analiștilor să identifice regiuni cu valori mai mari sau mai mici, disparități în indicatorii socio-economici sau hotspot-uri regionale. Astfel de informații pot ajuta la țintirea unor regiuni specifice pentru intervenții sau alocarea de resurse.

A map of the united states

Description automatically generated

Harta de mai sus evidențiază diferențe semnificative de venituri între state. De exemplu, state precum California, Texas și New York au venituri semnificativ mai mari în comparație cu state mai mici precum Rhode Island, Delaware sau Maine. Aceasta indică variații ale potențialului pieței și ale activității economice în diferite regiuni.

Harta dezvăluie, de asemenea, state cu venituri relativ mai mici, care pot prezenta oportunități de creștere neexploatate. De exemplu, state precum Oregon, Nevada și Idaho au venituri sub 2 milioane de dolari. Aceste regiuni ar putea fi vizate pentru dezvoltarea pieței, extinderea produselor sau inițiative de marketing pentru a capta o cotă de piață mai mare și a crește veniturile.

## **BAR CHART**

Bar chart sau diagrama cu bare este un tip de diagramă care utilizează bare dreptunghiulare pentru a reprezenta valorile datelor. Fiecare bară corespunde unei anumite categorii sau variabile, iar lungimea sau înălțimea barei reprezintă valoarea asociată acelei categorii. Diagramele cu bare sunt utilizate în mod obișnuit în vizualizarea datelor pentru a compara și afișa date categoriale sau valori discrete. Importanța diagramelor cu bare în analiza Big Data constă în capacitatea lor de a reprezenta vizual și de a comunica eficient informațiile.

A picture containing text, screenshot, design

Description automatically generatedAcest tip de diagramă facilitează compararea datelor, analiza distribuției, explorarea și prezentarea datelor. Prin utilizarea sa, analiștii identifică modele, iau decizii informate pe baza reprezentării vizuale a datelor.

Diagrama cu bare reprezintă analiza veniturilor bazată pe vârstă. Fiecare bară din grafic corespunde unui anumit grup de vârstă, iar lungimea barei reprezintă venitul asociat cu acel grup de vârstă.

Pe baza acestei interpretări, putem observa că grupa de vârstă 30-40 de ani generează cele mai mari venituri, urmată de grupa de vârstă de 60-70 de ani. Aceste informații pot fi valoroase pentru luarea deciziilor, cum ar fi direcționarea eforturilor de marketing către grupul de vârstă cu cel mai mare potențial de venituri obținute sau adaptarea produselor/serviciilor la anumite segmente de vârstă pentru a maximiza profitabilitatea. În plus, diagrama cu bare ajută la vizualizarea distribuției veniturilor obținute pe diferite grupe de vârstă, oferind informații despre comportamentul clienților pe baza datelor demografice de vârstă.

## **DONUT CHART**

Donut chart sau diagrama gogoașă este o diagramă circulară, un grafic statistic circular și este folosit pentru a vizualiza proporțiile sau procentele datelor într-o manieră clară și concisă. Inelul exterior al diagramei reprezintă setul total de date, în timp ce gaura interioară poate fi lăsată goală sau utilizată pentru a afișa informații suplimentare sau subcategorii.

Importanța graficelor donut în analiza Big Data constă în capacitatea lor de a transmite proporțiile și contribuțiile relative ale diferitelor categorii de date. Mărimea fiecărei felii reprezintă proporția sau procentul cu care contribuie la întreg, ceea ce face ușoară compararea și identificarea categoriilor semnificative. Forma circulară a diagramei gogoși permite o reprezentare clară a segmentelor de date. Simplifică informațiile complexe și permite spectatorilor să înțeleagă rapid imaginea de ansamblu. Diagramele donut pot fi folosite pentru a compara mai multe seturi de date sau categorii într-o singură diagramă. Prin utilizarea diferitelor culori sau modele, devine ușor să diferențiezi și să compari dimensiunile diferitelor categorii, evidențiind variațiile și tendințele.

A picture containing text, screenshot, font, diagram

Description automatically generated

Graficul reprezintă distribuția vânzărilor în diferite regiuni. Fiecare regiune este reprezentată de un segment, iar dimensiunea fiecărui segment corespunde proporției de vânzări atribuite acelei regiuni.

Regiunile de Sud și Vestul Mijlociu apar ca piețe cheie, cu cele mai mari contribuții la vânzări. Aceste informații pot fi utile pentru luarea unor decizii, cum ar fi concentrarea vânzărilor asupra regiunilor cu cel mai mare potențial de vânzări, identificarea zonelor de extindere sau îmbunătățirea și alocarea strategică a resurselor pentru a maximiza vânzările între regiuni. În timp ce aceste regiuni domină distribuția vânzărilor, regiunile de Nord-Est și Vest dețin în continuare cote semnificative. Înțelegerea nuanțelor fiecărei regiuni, inclusiv datele demografice și tendințele pieței, poate ghida dezvoltarea de strategii direcționate.

## **BUTTERFLY PLOT**

Butterfly plot sau diagrama fluture este o variație specializată a unei diagrame cu bare, care este utilizată în principal pentru a compara două seturi de date sau categorii înrudite una lângă cealaltă. Este format din două seturi de bare, poziționate orizontal pe diagramă, cu o linie de bază comună în centru. Barele din partea stângă reprezintă o categorie sau set de date, în timp ce barele din partea dreaptă reprezintă cealaltă categorie sau set de date.

Graficul fluture este util în situațiile în care se dorește sublinierea comparației și a contrastului dintre două seturi de date sau categorii. Permite compararea vizuală a valorilor și distribuțiilor celor două seturi într-o manieră compactă și intuitivă. Acest tip de diagramă este adesea folosit în domenii precum finanțe, economie și cercetare științifică pentru a evidenția efectele înainte și după o anumită intervenție sau compararea performanței a două entități.

În analiza Big Data, diagrama fluture poate fi valoroasă atunci când se examinează impactul diferitelor variabile sau factori asupra unui rezultat dat. Reprezentând vizual datele în acest format, se pot identifica rapid modele, tendințe și diferențe între cele două seturi.

A picture containing text, line, plot, number

Description automatically generated

Diagrama de mai sus reprezintă analiza veniturilor bazată pe gen. În mijloc sunt enumerate categoriile, cu veniturile genului feminin afișate în partea stângă și veniturile genului masculin în partea dreaptă. Lungimile sau înălțimile barelor din fiecare segment indică mărimea sau valoarea veniturilor corespunzătoare. Graficul fluture ne permite să comparăm veniturile generate de diferite categorii pe segmente de gen.

Graficul fluture arată o distribuție relativ echilibrată a veniturilor între bărbați și femei în diferitele categorii. Venitul total generat de femei și bărbați pare a fi similar, indicând un model de cheltuieli de consum relativ echitabil. Categoria „Telefoane mobile și tablete” se remarcă drept cel mai mare generator de venituri atât pentru femei, cât și pentru bărbați. Acest lucru sugerează o cerere puternică de dispozitive mobile și tablete în rândul ambelor sexe, subliniind importanța acestei categorii în stimularea vânzărilor generale. Categoriile „Cărți” și „Școală și educație” par să genereze venituri minime atât pentru femei, cât și pentru bărbați. Această observație sugerează necesitatea de a explora strategii de creștere a angajamentului și a cheltuielilor în aceste domenii, potențial prin promoții direcționate, parteneriate sau oferte de produse inovatoare.

## **SCATTER PLOT**

Scatter plot sau grafic de dispersie este un tip de vizualizare a datelor care afișează relația dintre două variabile numerice. Este creat prin trasarea punctelor de date individuale pe un sistem de coordonate carteziene, cu o variabilă reprezentată pe axa X și cealaltă variabilă reprezentată pe axa Y. Fiecare punct de date este reprezentat printr-un punct sau un marker la coordonatele corespunzătoare. Graficul de dispersie permite o examinare vizuală a relației dintre cele două variabile, dezvăluind modele, tendințe sau corelații care pot exista. Prin analizarea distribuției punctelor de date de pe diagramă, devine posibil să se obțină informații și să se ia decizii bazate pe date.

În analiza Big Data, diagramele de dispersie joacă un rol important în mai multe moduri, ca de exemplu:

* Analiza relațiilor: diagramele de dispersie permit explorarea relațiilor dintre două variabile. Ele pot ajuta la identificarea corelațiilor pozitive (ambele variabile cresc sau scad împreună), corelații negative (o variabilă crește în timp ce cealaltă scade) sau absența oricărei corelații. Aceste informații sunt cruciale pentru înțelegerea dependențelor și asocierilor dintre variabilele din seturile mari de date.
* Detectarea valorii aberante: diagramele de dispersie pot identifica valorile aberante, care sunt puncte de date care se abat semnificativ de la tiparul general. Valorile aberante pot reprezenta anomalii sau observații neobișnuite care necesită investigații suplimentare. Detectarea și analiza valorilor aberante poate oferi informații valoroase asupra problemelor de calitate a datelor, cazuri excepționale sau modele ascunse.

A picture containing text, screenshot, line, plot

Description automatically generated

Pe baza diagramei de dispersie care arată corelația Cantitate - Discount, putem observa un model clar care apare. Pe măsură ce procentul de reducere crește, crește și cantitatea de comenzi. Aceasta indică o corelație pozitivă între cele două variabile.

Corelația pozitivă sugerează că există o relație între discountul oferit la un produs și cantitatea de comenzi plasate. Atunci când un produs are un procent de reducere mai mare, acesta tinde să atragă mai mulți clienți, rezultând o cantitate mai mare de comenzi. Acest model se aliniază cu intuiția comună că clienții sunt mai susceptibili să cumpere un produs atunci când acesta este oferit la un preț redus. Graficul de dispersie ne permite să confirmăm vizual această relație și să identificăm puterea corelației.

# **DASHBOARD**

Tableau pune la dispoziție crearea dashboard-urilor, oferind o modalitate atrăgătoare și interactivă de a prezenta datele. Acest lucru îmbunătățește înțelegerea datelor și permite utilizatorilor să identifice rapid modele, tendințe și valori.

Acest Dashboard permite utilizatorilor să interacționeze cu datele aplicând filtre, selectând dimensiuni specifice și analizând detalii. Se pot explora diferite perspective ale datelor, analizarea variabilelor și obținerea unor informații mai profunde. Caracteristicile interactive îmbunătățesc capacitatea de explorare.

A screenshot of a computer screen

Description automatically generated with low confidence

Dashboard-ul prezentat conține o serie de caracteristici care îl fac util și valoros pentru analiza datelor. Dropdown-ul dinamic reprezintă un element interactiv, care permite utilizatorului să selecteze o categorie specificată. Odată selectată o categorie, graficele și datele prezentate în dashboard se vor actualiza în consecință, furnizând informații relevante și specifice categoriei selectate. Acesta facilitează analiza în profunzime a datelor în funcție de criteriile dorite.

În partea de sus a dashboard-ului, există un label care indică veniturile totale. Label-ul oferă o perspectivă generală asupra performanței globale. Prin furnizarea unei vizualizări agregate și detaliate a datelor, dashboard-ul facilitează procesul de luare a deciziilor bazate pe date.

# **CONCLUZII**

În concluzie, tehnicile de vizualizare a datelor joacă un rol esențial în analiza Big Data și au o importanță considerabilă în procesul de extragere a informațiilor și luarea deciziilor. Aceste tehnici permit transformarea datelor complexe și voluminoase în reprezentări vizuale intuitive și ușor de înțeles, permițând utilizatorilor să exploreze, să identifice modele și tendințe și să extragă cunoștințe relevante.

Una dintre cele mai mari provocări o reprezintă volumul și diversitatea informațiilor disponibile. Cu ajutorul tehnicilor de vizualizare a datelor, aceste informații pot fi structurate și prezentate într-o manieră coerentă, astfel încât să ofere o imagine de ansamblu clară și accesibilă. Graficele, diagramele și alte forme de vizualizare, permit identificarea rapidă a unei conexiuni, anomalii, ceea ce facilitează procesul de înțelegere. Totodată, tehnicile de vizualizare a datelor permit comunicarea eficientă a rezultatelor analizei Big Data. Prin prezentarea informațiilor sub formă vizuală, se elimină barierele lingvistice și tehnice, facilitând comunicarea între profesioniștii din diferite domenii și toate nivelurile organizaționale. Prin împărtășirea rezultatelor analizei sub forma unor grafice sau diagrame ușor de interpretat, utilizatorii pot asimila informațiile și pot colabora mai eficient pentru adoptarea unor strategii.

Astfel, tehnicile de vizualizare a datelor sunt esențiale în analiza Big Data, deoarece oferă o abordare practică pentru a gestiona și a înțelege seturi de date imense.

# **BIBLIOGRAFIE**

1. *17 Important Data Visualization Techniques* | HBS Online. (2019, September 17). Business Insights Blog. <https://online.hbs.edu/blog/post/data-visualization-techniques>

Accesat la: 12.05.2023

1. Gill, N. S. (2022, December 16). *Big Data Visualization Techniques and Tools* | Quick Guide. XenonStack. <https://www.xenonstack.com/blog/big-data-visualization-techniques>

Accesat la: 12.05.2023

1. Writer, R. C. (2022). *Big data visualization: what it is, techniques and best tools*. Rock Content. <https://rockcontent.com/blog/big-data-visualization/>

Accesat la: 15.05.2023

1. *Sales Dataset*. (2023, May 5). Kaggle. <https://www.kaggle.com/datasets/miketaylor123/sales-data>

Accesat la: 15.05.2023

1. Scifres. (2021, December 8). *Best Data and Big Data Visualization Techniques* - Sciforce - Medium. Medium. <https://medium.com/sciforce/best-data-and-big-data-visualization-techniques-e07b897751dd>

Accesat la: 16.05.2023

1. Beker, A. (2022, July 19). *Big Data Visualization: Value It Brings and Techniques It Requires*. Science Soft. <https://www.scnsoft.com/blog/big-data-visualization-techniques>

Accesat la: 16.05.2023