MINISTERUL EDUCAȚIEI ȘI CERCETĂRII

UNIVERSITATEA DE STAT DIN MOLDOVA

FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ

DEPARTAMENTUL INFORMATICĂ

RAPORT

la disciplina “Analiza și Vizualizarea Datelor”

Autorul: Cojocari Adriana, grupa IASD2501

Conducător ştiinţific: Poiata Anatol, lector universitar

CHIŞINĂU 2025

1. Rezumat

Această lucrare are ca obiectiv aplicarea tehnicilor fundamentale de analiză și vizualizare a datelor pe un set de date complex din domeniul vitivinicol. Proiectul urmărește explorarea relațiilor dintre caracteristicile numerice, categorice și textuale ale vinurilor, cu accent pe identificarea corelațiilor dintre preț, punctaj, conținutul de alcool și descrierile textuale. Metodologia include etape de preprocesare a datelor, analiză statistică descriptivă, analiză de text și generarea de vizualizări avansate, inclusiv reprezentări interactive. Rezultatele obținute evidențiază tipare relevante în date, confirmă existența unor corelații între variabilele analizate și demonstrează utilitatea vizualizării datelor în sprijinul procesului decizional.

2. Introducere

Analiza și vizualizarea datelor reprezintă componente esențiale în extragerea de informații utile din seturi mari de date, un proces din ce în ce mai valoros în epoca digitală. În domeniul vitivinicol, aceste tehnici permit investigarea relațiilor dintre caracteristicile vinurilor - cum ar fi prețul, punctajul acordat de critici, conținutul de alcool sau descrieri textuale - pentru a descoperi tipare ascunse și corelații semnificative, ceea ce poate influența atât deciziile consumatorilor, cât și strategiile producătorilor.

Problema abordată în această lucrare este determinarea modului în care diverse atribute ale vinurilor se corelează între ele și modul în care aceste relații pot fi comunicate eficient prin intermediul vizualizărilor. În particular, se investighează dacă prețul unei sticle de vin este asociat cu calitatea măsurată prin punctaj, dar și impactul altor variabile importante, cum ar fi conținutul de alcool sau descrierile senzoriale ale vinurilor.

Motivația acestui proiect este dublă: pe de o parte, oferă o aplicare practică a tehnicilor teoretice învățate, iar pe de altă parte, rezultatele obținute pot sprijini înțelegerea aprofundată a factorilor care influențează percepția calității vinului. Vizualizările generate, cum ar fi histograme, grafice de dispersie și heatmap-uri, facilitează interpretarea relațiilor complexe și permit evidențierea tendințelor cheie din date.

Intrarea în algoritmul de analiză este reprezentată de un set de date despre vinuri ce include coloane numerice (price, points, alcohol), categorice (country, category, variety, region_1/2) și textuale (description), în timp ce ieșirile constau în statistici descriptive, analize de corelație, vizualizări grafice și o aplicație interactivă de explorare a datelor.

3. Lucrări conexe

În domeniul analizelor asupra vinurilor utilizând tehnici de date, literatura existentă poate fi grupată în trei categorii principale:

a) Analiza statistică și modelare a review-urilor de vinuri.

Studiile din această categorie investighează seturi mari de recenzii pentru a compara consistența evaluatorilor sau pentru a cuantifica caracteristicile relevante. De exemplu, Wineinformatics: A Quantitative Analysis of Wine Reviewers [1] analizează peste 100.000 de recenzii pentru a compara consistența criticilor faimoși și folosește clasificatoare pentru evaluare.

b) Analiza descriptivă și vizualizarea datelor din review-uri.

Proiecte precum Analysis and visualization of WineEnthusiast wine reviews [2] oferă exemple practice de explorare a distribuțiilor de puncte, prețuri și varietăți, precum și a celor mai frecvente cuvinte în descrieri pentru un set de date de mii de vinuri. Un alt proiect open-source, WineReviewAnalysis [3], explorează relațiile dintre preț și punctaj și creează vizualizări standard pentru un set de ~150.000 de review-uri.

c) Analiza calității vinului pe baza caracteristicilor fizico-chimice.

Deși diferită în structura datelor - concentrându-se pe teste de laborator mai degrabă decât pe descrieri și prețuri - această categorie oferă insight-uri despre modelarea calității vinului. Proiecte precum Red Wine Quality Analysis [4] și alte analize folosind datele UCI explorează factori precum aciditatea sau pH-ul în corelație cu scorul de calitate.

Puncte tari și slabe:

- Lucrările din prima categorie tratează corect dimensiunea și complexitatea datelor textuale, dar se concentrează mai puțin pe vizualizări intuitive pentru utilizatori non-tehnici.
- Analizele din a doua categorie sunt aplicabile direct proiectului, cu vizualizări clare, dar deseori rămân la nivel explorator și nu includ aplicații interactive.
- Modelele bazate pe caracteristici fizico-chimice oferă predicții solide, însă nu exploatează textul descriptiv sau variabilele economice.

Diferențierea soluției actuale: Proiectul realizat combină analiza statistică, vizualizarea avansată și analiza textului într-un singur cadru, incluzând și o aplicație interactive în Streamlit pentru filtrare și căutare în descrieri, ceea ce extinde utilitatea soluțiilor prezentate în literatura de specialitate. Această abordare integrată pune în evidență atât corelațiile tradiționale între variabile numerice, cât și impactul textual, oferind o imagine completă asupra datelor despre vinuri.

4. Set de date și caracteristici

Setul de date utilizat în cadrul acestui proiect conține informații despre vinuri evaluate de experți, incluzând atât atribute numerice, cât și categorice și textuale. După procesul complet de curățare și preprocesare, setul de date final conține 57.196 de înregistrări, fiecare corespunzând unui vin unic. Fiecare observație este descrisă prin variabile precum prețul (price), punctajul acordat (points), conținutul de alcool (alcohol), anul recoltei (vintage), țara de origine (country), categoria vinului (category), soiul de struguri (variety), precum și descrierea textuală a vinului (description).

Înregistrări brute:

```
country,description,designation,points,price,province,region_1,region_2,variety,winery,title,vintage,alcohol,category
Spain,"Consistent with past vintages, this crianza gets it right without setting off any balloons. It starts with raspberry and cassis aroma
US,"The Indelicato family's vineyard is basically the entire appellation. This hearty blend of Syrah, Grenache, Petit Verdot and Viognier sh
```

După preprocesare completă (wine_clean_final.csv):

```
country,description,designation,points,price,province,region_1,region_2,variety,winery,title,vintage,alcohol,category,price_quality_ratio
spain,consistent past vintages crianza gets right setting balloons starts raspberry cassis backs slightly citrusy red rawness heat tamed mat
usa,indelicato family vineyard basically entire appellation hearty blend syrah grenache petit verdot viognier shows driven grape juice grape
```

Preprocesarea datelor:

Preprocesarea datelor a reprezentat o etapă esențială pentru asigurarea calității analizei. Inițial, au fost identificate valorile lipsă pentru fiecare coloană. Valorile lipsă din variabilele numerice au fost completate folosind mediana, metodă robustă la valori extreme. Ulterior, au fost eliminate înregistrările duplicate.

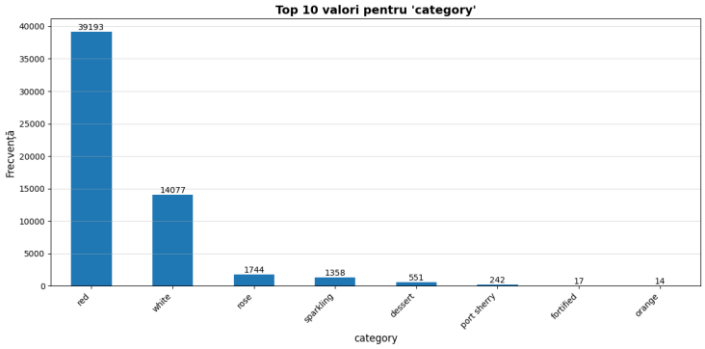
Coloanele numerice (price, points, vintage, alcohol) au fost convertite explicit la tip numeric. Pentru variabila alcohol, au fost corectate valori anormale rezultate din erori de scalare, prin normalizarea acestora (de exemplu împărțire la 10 sau 100, în funcție de caz). De asemenea, au fost eliminate valorile invalide pentru anul recoltei (vintage > 2025).

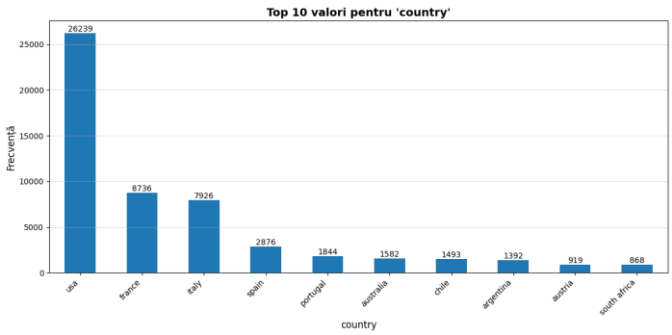
Datele categorice și textuale au fost standardizate prin:

- conversie la litere mici,
- eliminarea caracterelor non-alfanumerice,
- eliminarea spațiilor multiple,
- uniformizarea denumirilor pentru țări (ex.: „us”, „united states” → „usa”).

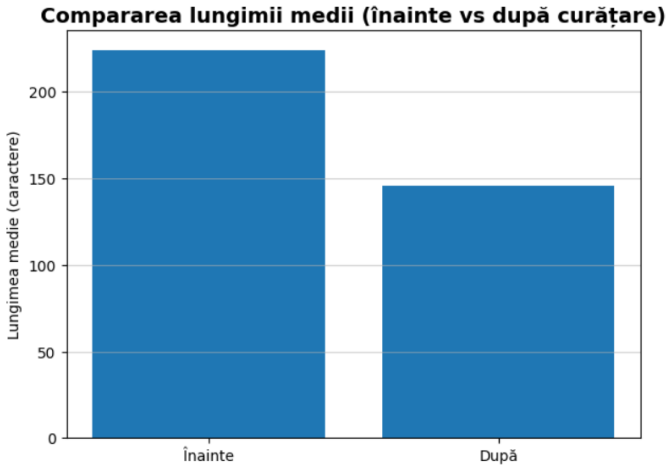
În plus, a fost creată o variabilă nouă, raportul preț/calitate (price_quality_ratio), calculată ca raport între preț și punctaj, pentru a facilita analiza eficienței economice a vinurilor.

Exemple de date:

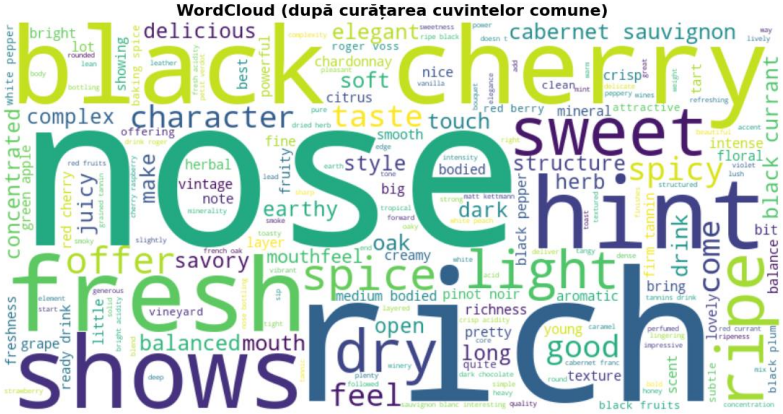




Pentru datele textuale, descrierile vinurilor au fost curățate prin eliminarea cuvintelor comune (stopwords generale și specifice domeniului). Lungimea medie a descrierilor a scăzut de la 224,31 caractere înainte de curățare la 145,96 caractere după curățare, indicând o reducere semnificativă a zgomotului textual.



Vizualizareaa celor mai frecvente cuvinte din descriere după eliminarea stopwords:



Analiza s-a concentrat pe:

- Caracteristici numerice: price, points, alcohol, vintage, precum și raportul preț/calitate, utilizate pentru statistici descriptive și analize de corelație.
- Caracteristici categorice: country, category, variety, utilizate pentru distribuții și vizualizări comparative.
- Caracteristici textuale: descrierile vinurilor, procesate prin curățare și analiză de frecvență a cuvintelor. Analiza textuală a fost realizată la nivel lexical, scopul fiind explorarea interpretabilă a conținutului textual și corelarea termenilor frecvenți cu prețul și punctajul.

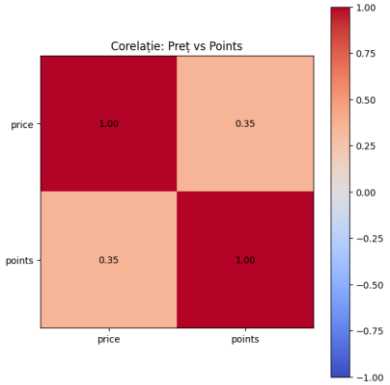
5. Experimente și rezultate

Scopul experimentelor a fost explorarea relațiilor dintre variabilele numerice, categorice și textuale ale datasetului de vinuri, precum și vizualizarea acestor corelații. Analiza a inclus:

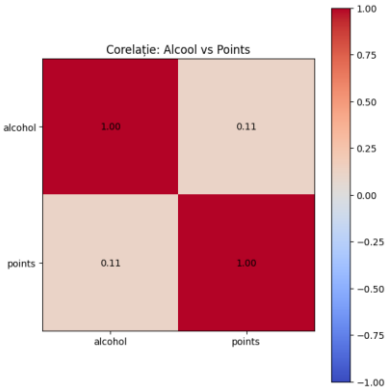
- Corelații numerice și între variabile categorice și preț/punctaj;
- Corelații între frecvența cuvintelor din descriere și preț/punctaj sau soiuri;
- Vizualizări avansate precum histogramă, bar plot, stacked bar chart și scatter plot interactiv.

Corelații numerice:

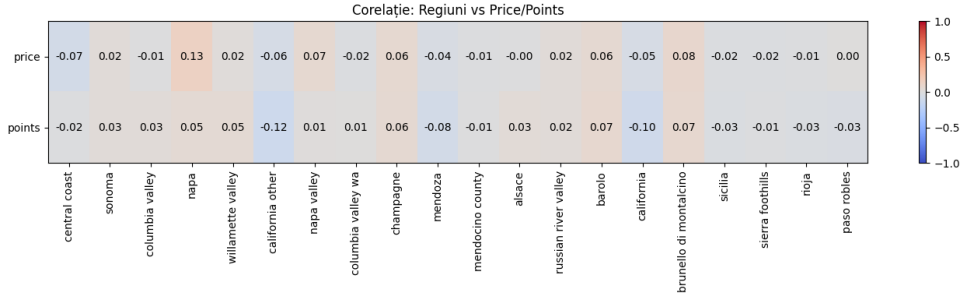
Price vs Points: Coeficientul Pearson: 0.35, indicând o corelație moderată pozitivă. Vinurile cu punctaj mai mare tind să aibă prețuri mai ridicate.



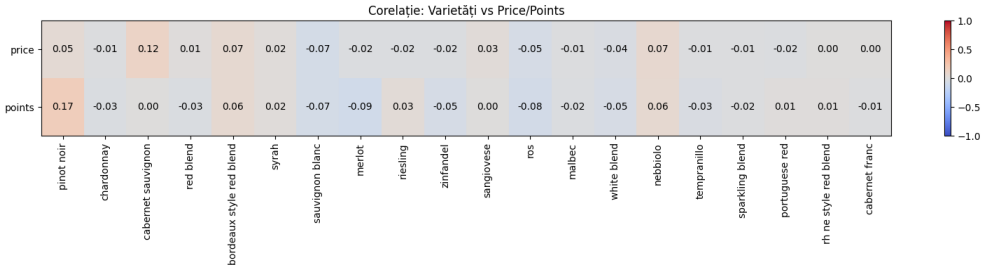
Alcohol vs Points: Coeficientul Pearson: 0.11, indicând o corelație slabă pozitivă. Conținutul de alcool nu determină semnificativ punctajul.



Regiune vs Price/Points: Heatmap pentru top 20 regiuni arată că anumite au vinuri cu prețuri și punctaje mai mari:

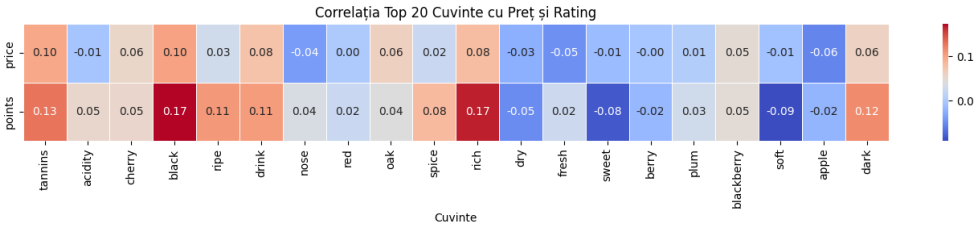


Varietăți vs Price/Points: Vinurile Pinot Noir și Chabernet Sauvignon apar mai frecvent printre cele mai scumpe și apreciate:

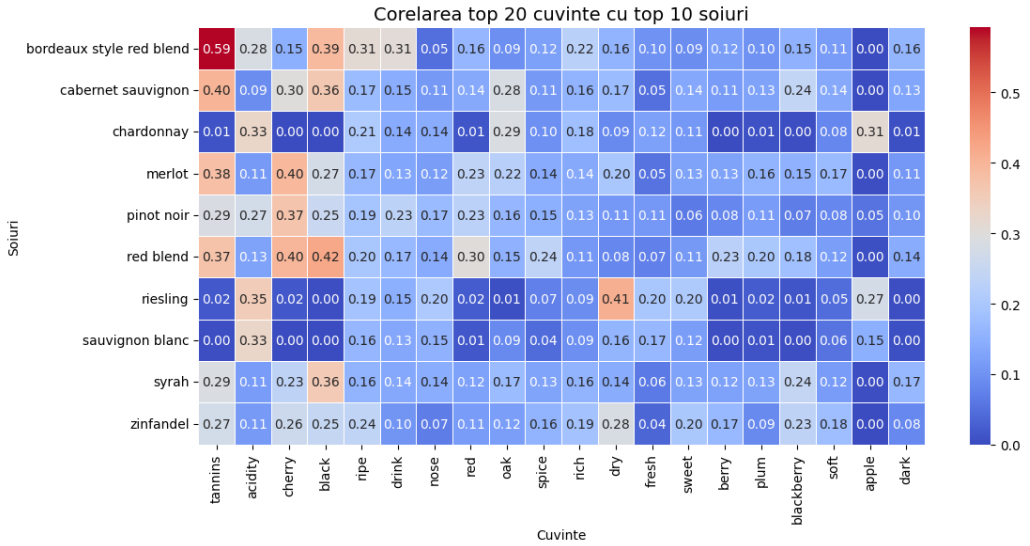


Analiza textului :

Corelarea cuvintelor cele mai frecvente cu prețul și ratingul: cuvintele „tannins”, „black” , „rich” , „dark”, au corelații pozitive cu punctajul și prețul, indicând că descrierile bogate în termeni de calitate corespund scorurilor și prețurilor mai mari.

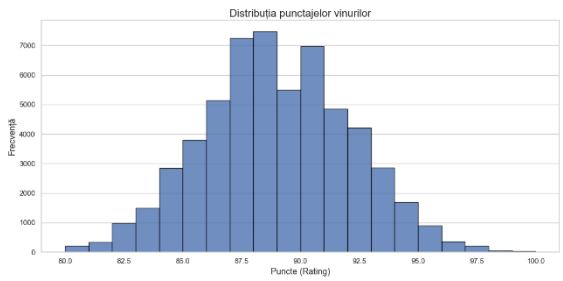


Corelarea top 20 cuvinte cu top 10 soiuri: Heatmap-ul prezintă frecvența medie a celor mai comune 20 de cuvinte din descrieri pentru fiecare dintre cele 10 soiuri de vin cele mai populare, evidențiind termenii descriptivi care apar cel mai des în asociere cu anumite soiuri.

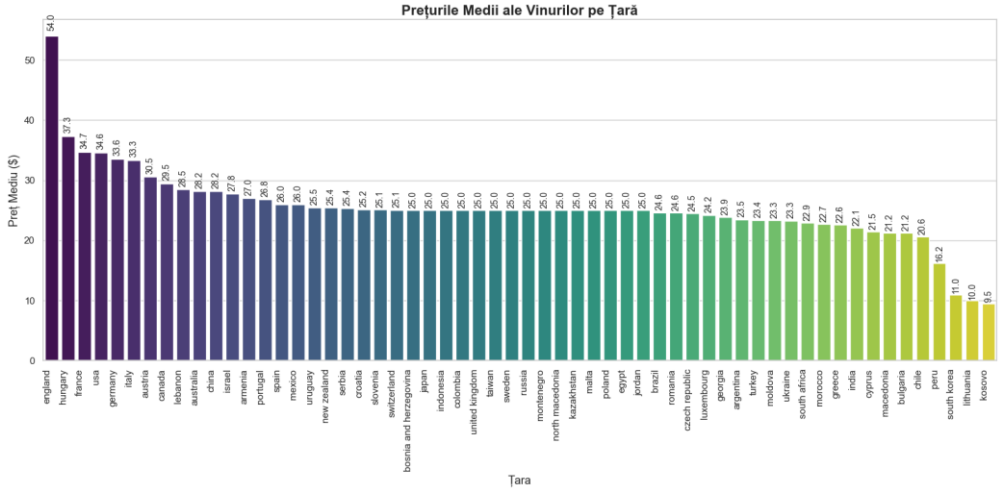


Vizualizări avansate:

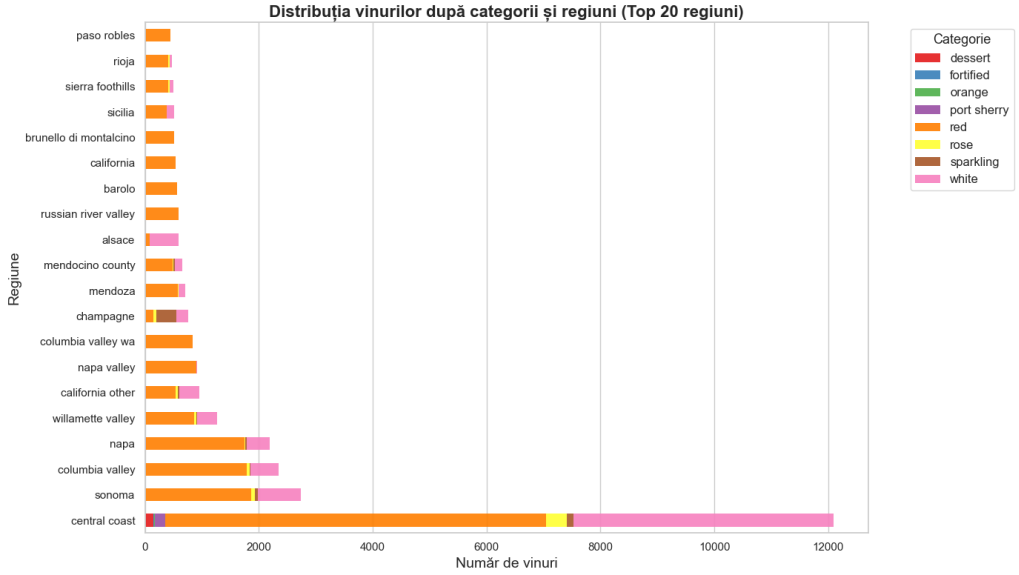
Distribuția punctajelor: histogramă arată că majoritatea vinurilor au punctaje între 86-91:



Prețuri medii pe țară: bar plot-ul evidențiază că Anglia are vinuri mai scumpe, fiind in top, urmată de Ungaria, Franța, USA, Germania și Italia



Distribuția vinurilor după categorii și regiuni: Acest grafic arată cum se distribuie diferitele categorii de vinuri (roșu, alb, rose, dessert, etc.) în cadrul celor mai importante 20 de regiuni, evidențiind cantitatea relativă a fiecărei categorii pe regiune.



6. Realizări și implementări

S-a realizat o aplicație web complet funcțională în Streamlit pentru analiza interactivă a vinurilor. Principalele funcționalități sunt:

- Încărcare și preprocesare automată a dataset-ului wine_clean_final.csv, inclusiv calculul raportului preț/calitate și vectorizare text.
- Sidebar cu filtre multiple: țară, categorie, interval preț, interval punctaj, raport preț/calitate, căutare textuală în descrieri.
- 5 tab-uri tematice cu peste 13 vizualizări diferite (corelații, heatmap-uri, scatter plot-uri, stacked bar chart-uri, distribuții și vizualizări interactive Plotly).
- Butoane de descărcare PDF pentru fiecare grafic.
- Interfață intuitivă și responsivă, cu explicații și legende pentru fiecare vizualizare.

Aplicația rulează local cu streamlit run app.py și toate graficele sunt reproduse exact ca în raport.

7. Concluzie și implementări viitoare

Proiectul a permis analiza detaliată a dataset-ului de vinuri, evidențiind relațiile dintre preț, punctaj, regiuni, varietăți și caracteristici lingvistice din descrieri. Algoritmul cel mai eficient în explorarea datelor a fost analiza corelațiilor combinate cu vizualizările interactive, care au oferit perspective clare asupra tipurilor de vinuri premium și a factorilor ce influențează calitatea percepută.

Direcții viitoare posibile:

• Extinderea dataset-ului cu date suplimentare (recenzii, anotări utilizatori, condiții climatice).
• Implementarea de modele predictive pentru recomandarea vinurilor și identificarea celor cu cel mai bun raport preț/calitate.
• Adăugarea de vizualizări și statistici comparative între ani sau regiuni.
• Crearea unui modul interactiv de raportare automată și export PDF/Excel pentru analiza profesională.

8. Referințe

[1] Wineinformatics: A Quantitative Analysis of Wine Reviewers , 2025. https://www.mdpi.com/2311-5637/4/4/82
[2] Analysis and Visualization of WineEnthusiast Wine Reviews, 2025. https://manuelenolli.github.io/wine-enthusiast-analysis/
[3] WineReviewAnalysis (GitHub) – https://github.com/gorbulus/WineReviewAnalysis
[4] Red Wine Quality Analysis (GitHub) – https://github.com/vikrantkakad/Red-Wine-Quality-Analysis
[5] Wine-Quality-Analyzation-by-R (GitHub) – https://github.com/gaallmin/Wine-Quality-Analyzation-by-R
[6] Pandas Development Team. pandas: powerful Python data analysis toolkit. https://pandas.pydata.org
[7] Matplotlib Development Team. Matplotlib: A 2D graphics environment. https://matplotlib.org
[8] Seaborn Development Team. seaborn: statistical data visualization. https://seaborn.pydata.org