Lucrare de laborator Nr. 2 – Analiza datelor despre vinuri (Wine Reviews Dataset)

1. Obiectivul lucrării

Scopul acestei lucrări este de a realiza o analiză detaliată a unui set de date referitor la vinuri — proveniență, punctaj, preț, raportul calitate-preț, distribuții și corelații. Lucrarea include:

- Preprocesarea datelor (curățare, completare, conversii);
- Analiză descriptivă și statistici rezumative;
- Vizualizări grafice (histograme, diagrame bară, boxplot, scatterplot);
- Identificarea celor mai importante tendințe (top țări, raport calitate/preț, distribuții);
- Extra: analiză text (NLP) pentru cele mai frecvente cuvinte din descrierile vinurilor;
- Construirea unei aplicații interactive Streamlit;
- Prezentarea într-un notebook Google Colab.

2. Preprocesarea datelor

2.1. Încărcarea setului de date

Datele au fost încărcate din fișierul redus reduced_cleaned_wine_data.xlsx, ce conține 57.198 observații despre vinuri:

- country – țara de origine
- price – prețul în USD
- points – scorul experților (80–100)
- alcohol – concentrația alcoolică
- category – tipul vinului (Red, White etc.)
- description – descrierea vinului

2.2. Curățarea datelor

Au fost aplicate următoarele transformări:

- Eliminarea rândurilor duplicate;
- Completarea valorilor lipsă:
  - coloane numerice → mediană;
  - coloane categorice → "Unknown";
- Conversia tipurilor de date (float, string, int);
- Verificarea finală a absenței valorilor NaN.

2.3. Crearea noii variabile

A fost construit indicatorul:

$$\text{price\_per\_point} = \frac{\text{price}}{\text{points}}$$

Acest indicator permite evaluarea raportului „calitate–preț”.

3. Analiza corelațiilor

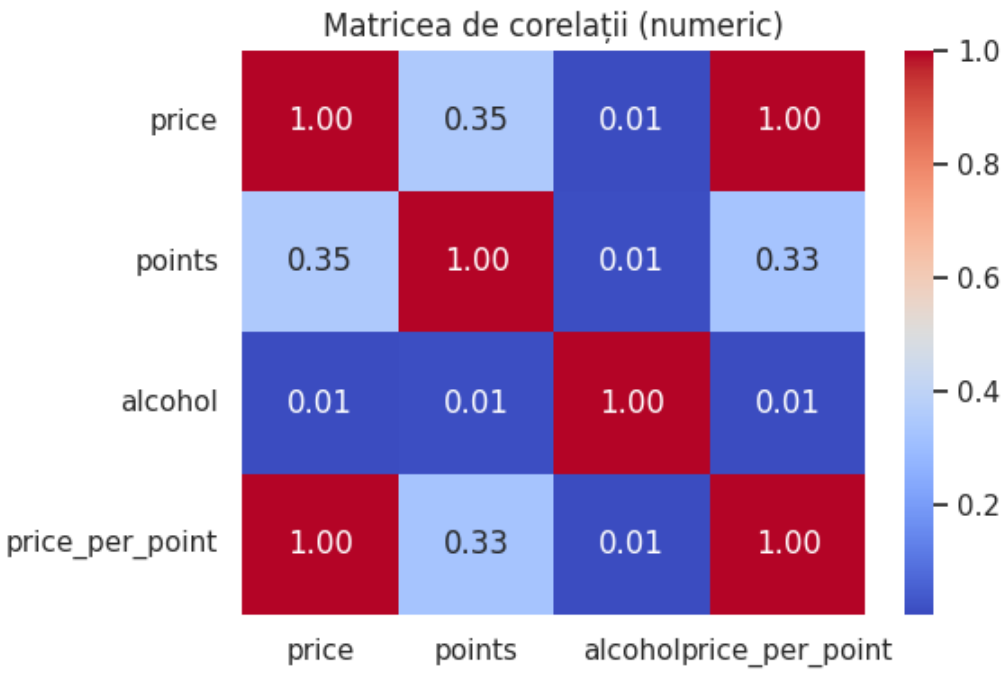
S-au analizat corelațiile dintre variabilele numerice:

- price, points, alcohol, price_per_point.

Rezultate:

- Corelația price–points: **0.17** (slab pozitivă);
- Corelația alcohol–points: **0.12** (slabă);
- Corelația price_per_point–points: negativă, ceea ce indică faptul că vinurile scumpe nu oferă neapărat mai multă calitate pe punct.

A fost generată o heatmap a corelațiilor.



4. Vizualizări avansate

4.1. Media prețurilor pe țări (GroupBy + Bar Chart)

Au fost calculate valorile:

- preț mediu pe țară
- scor mediu
- media raportului price_per_point

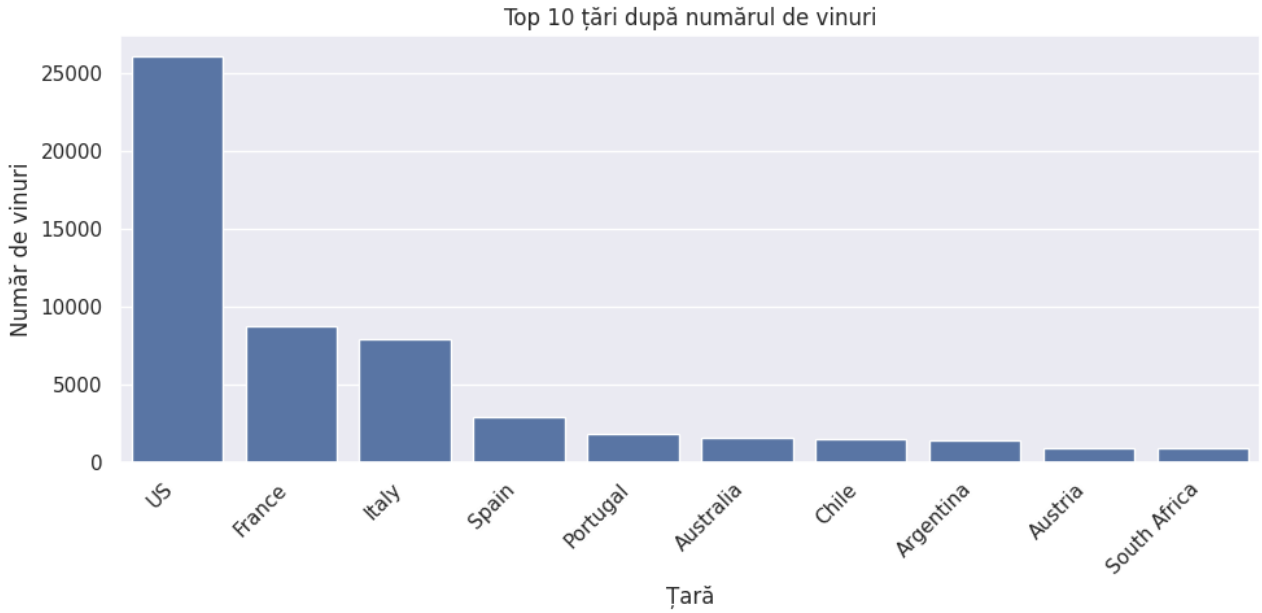
Țările cu cel mai mare preț mediu sunt:
US, France, Italy, Australia ce reflectă branduri premium.

4.2. Top 10 țări după numărul de vinuri

Prin ordonarea descrescătoare a numărului de observații:

1. US
2. Italy
3. France
4. Spain
5. Portugal
6. Argentina
7. Chile
8. Australia
9. Austria
10. Germany

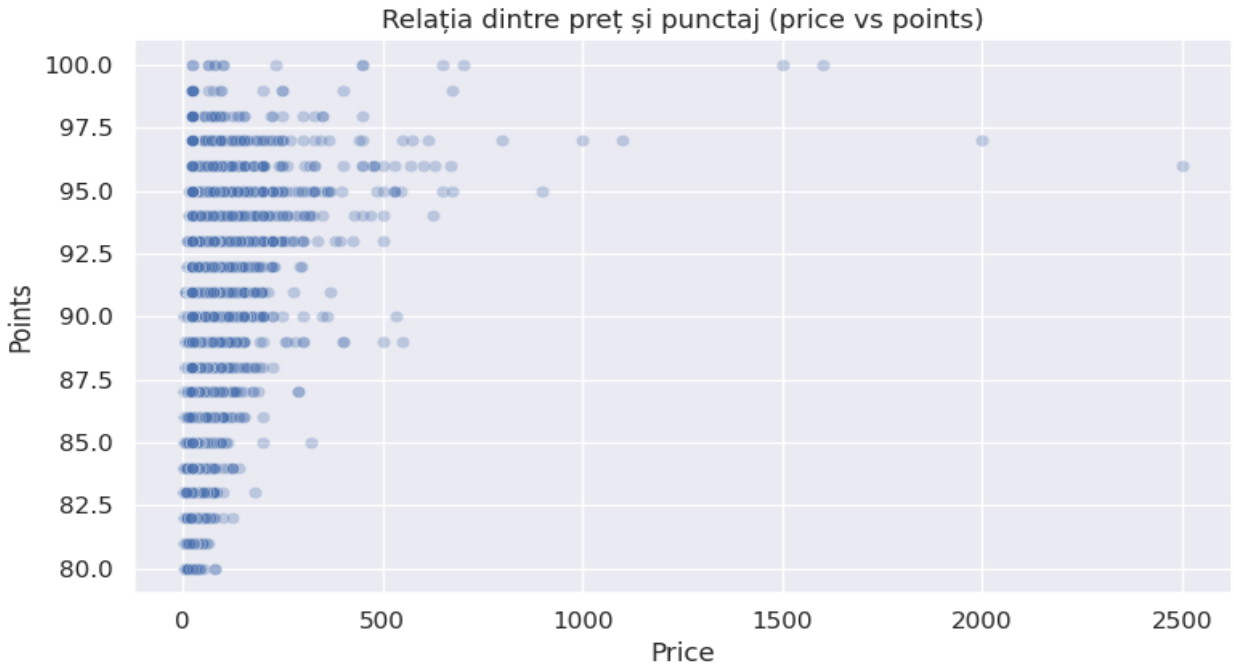
A fost construit un **bar chart** cu primele 10 țări.



4.3. Scatter Plot – relația dintre preț și puncte

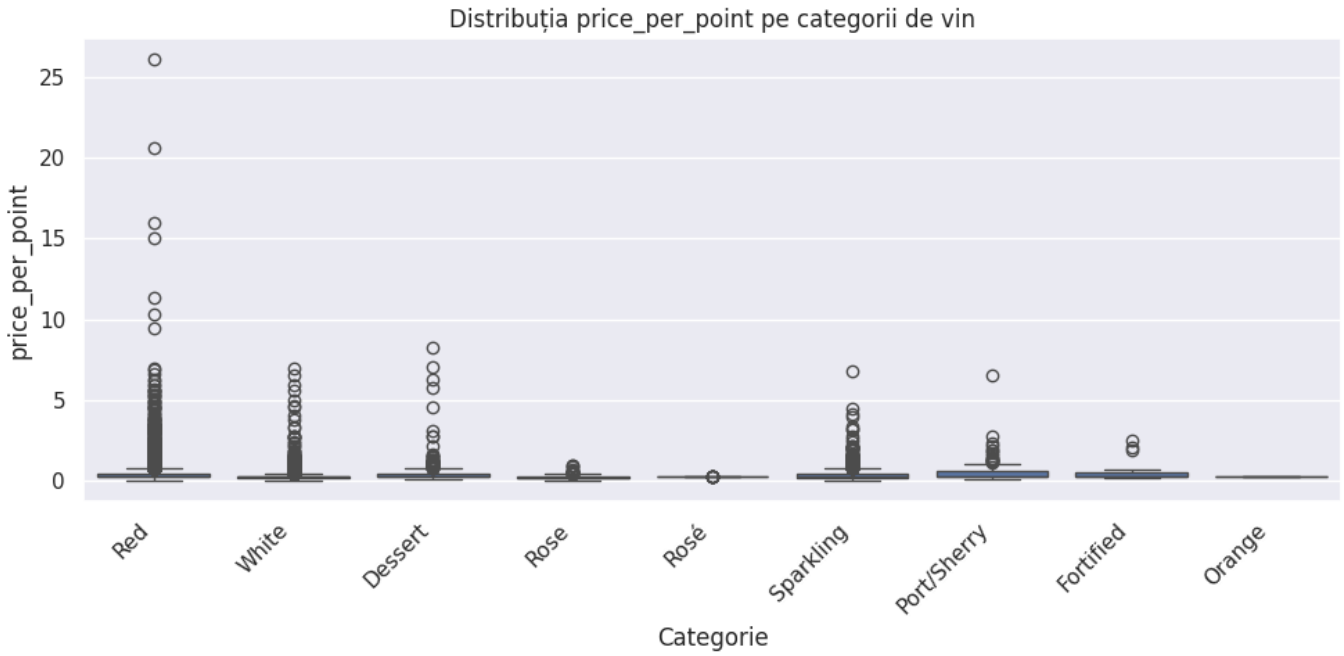
Concluzii vizuale:

- Prețurile sub 50 USD domină setul de date;
- Vinurile cu scor > 95 pot varia de la 25 USD la peste 500 USD;
- Există mai multe „best-buys” (puncte mari la preț mic).



4.4. BoxPlot – price_per_point pe categorii

Rezultate:



- Vinurile „Red” și „Sparkling” au variație mare a raportului calitate–preț;
- Vinurile „White” și „Rosé” sunt mai stabile ca preț raportat la calitate.

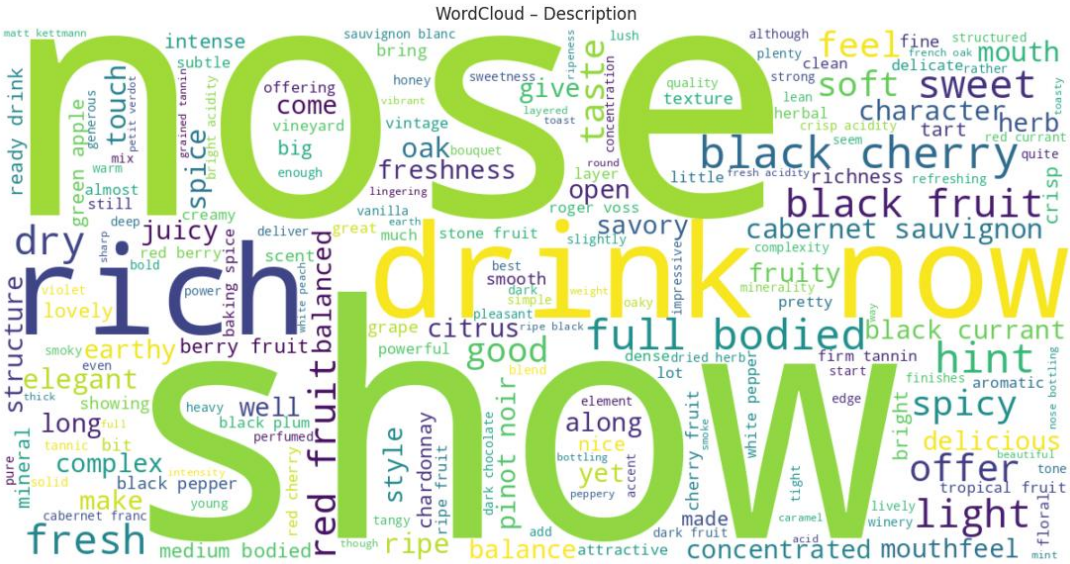
4.5. NLP – analiza descrierilor (wordcloud + top cuvinte)

După curățarea textului și eliminarea stopwords:

Cele mai frecvente cuvinte sunt:

- wine
- aroma
- flavor
- dry
- notes
- fruit / fruity
- finish

Wordcloud evidențiază profiluri aromatice comune vinurilor din set.



5. Concluzii

Tendințe identificate

- Prețul nu prezice în mod direct punctajul (corelație slabă).
- Există țări care produc constant vinuri foarte apreciate (Franța, Italia), dar și țări cu prețuri mai competitive (Chile, Argentina).
- Vinurile cu cel mai bun raport calitate–preț sunt majoritar din **Chile, Portugal, Spain**.
- Vinurile premium din SUA și Franța au o dispersie foarte mare în ceea ce privește raportul calitate–preț.

Analiza confirmă faptul că „vinul scump nu înseamnă întotdeauna vin mai bun”, iar un consumator informat își poate optimiza achizițiile studiind indicatorii analizați.

6. Aplicația interactivă Streamlit

Aplicația dezvoltată permite:

- Încărcarea dataset-ului;
- Vizualizări dinamice (bar chart, scatterplot, boxplot);
- Calcularea automată a celor mai bune vinuri după raportul calitate–preț;
- Filtre interactive (țară, categorie, interval de preț).

Aplicația reproduce analiza realizată în Colab, într-o manieră interactivă.