# Task: Analiza distribuției datelor din magazinul online

**Curs: Data Analysis and Processing using Python**
**Modul: Analiza distribuției datelor**

## Scopul sarcinii:

Dezvoltați capacitatea de a identifica și interpreta caracteristicile de distribuție a datelor într-un set de date real prin calculul măsurilor de variabilitate, uniformitate și asociere între atributele numerice ale produsului.

## Contextul sarcinii:

Alex vrea să planifice inventarul mai eficient și să îmbunătățească procesele de achiziție în comerțul online. Pentru a face acest lucru, trebuie să înțeleagă mai detaliat cum sunt distribuite anumite atribute ale produsului - cum ar fi evaluările, prețurile și numărul de unități vândute. Sarcinavoastră este să îl ajutați să analizeze distribuția acestor date folosind metode statistice standard.

## Sarcina:

Vom utiliza același set de date pe care Alex l-a extras din sistemul informațional al magazinului online. Îl puteți descărca de aici:

online_store_data.csv

Sarcina voastră este să scrieți un script Python care să răspundă la următoarele întrebări:

**1. Care este diferența dintre cel mai bine și cel mai slab evaluat televizor?**

- Utilizați intervalul (range = max - min)

**2. În ce interval de preț se află cel mai mare număr de telefoane mobile vândute?**

- Calculați intervalul intercuartil (IQR)

**3. Care sunt cele 5 branduri cu cele mai uniforme evaluări?**

- Calculați abaterea standard a evaluărilor

**4. Depinde numărul de evaluări de numărul de unități vândute? Mai multe unități vândute înseamnă mai multe evaluări sau invers?**

- Împărțiți produsele în cuartile în funcție de numărul de evaluări

## Ghid pentru rezolvarea sarcinii:

Mai jos sunt pașii care vă pot ajuta să îndepliniți sarcinile:

**1. Care este diferența dintre cel mai bine și cel mai slab evaluat televizor?**

- Pentru a rezolva această sarcină, mai întâi trebuie să filtrați setul inițial de date după categorie, astfel încât să rămână doar produsele care sunt în categoria „TVs”. Apoi, trebuie să găsiți valorile minime și maxime în coloana cu evaluări (rating). Diferența dintre valoarea minimă și cea maximă va reprezenta intervalul, adică exact ceea ce se cere în această sarcină.

**2. În ce interval de preț se află numărul tipic de telefoane mobile vândute?**

- Pentru a rezolva această sarcină, trebuie mai întâi să filtrați setul de

date al produselor după categorie, astfel încât să rămână doar produsele din categoria „Smartphones”. După aceea, trebuie să găsiți valorile care reprezintă prima și a treia cuartilă. Acesta este intervalul intercuartil și, de asemenea, intervalul de preț în care se află numărul tipic de telefoane mobile vândute.

**3. Care sunt cele 5 branduri cu cele mai uniforme evaluări?**

- Pentru a găsi cele 5 branduri cu cele mai uniforme evaluări, trebuie să utilizați abaterea standard. Mai întâi, trebuie să grupați produsele după branduri. Apoi, trebuie să calculați abaterea standard pentru fiecare grup. La final, trebuie să sortați setul de date obținut în funcție de valorile abaterii standard, în ordine crescătoare, și să selectați primele 5 rânduri. Acestea vor fi cele 5 branduri cu cele mai uniforme evaluări.

**4. Depinde numărul de evaluări de numărul de unități vândute? Mai multe unități vândute înseamnă mai multe evaluări sau invers?**

- Este clar că produsele cu un număr mai mare de vânzări au, de obicei, un număr mai mare de evaluări. Un produs care nu are vânzări nu poate avea evaluări. În această sarcină, trebuie să dovediți acest lucru. Pentru a face acest lucru, trebuie să utilizați cuartilele pentru a împărți setul de date în 4 părți. Mai întâi, trebuie să găsiți prima, a doua și a treia cuartilă pentru coloana „`num_of_ratings`”. Apoi, fiecărui rând îi va fi atribuită una dintre valorile „1st quartile”, „2nd quartile”, „3rd quartile”, „4th quartile”, în funcție de numărul de evaluări (valoarea din coloana num_of_ratings). După atribuirea acestor valori, le puteți folosi pentru a grupa datele. Pentru grupurile obținute, trebuie să însumați valorile din coloana numărului total de unități vândute (coloana `quantity_sold`). Rezultatul va arăta că produsele cu un număr mai mic de unități vândute au mai puține evaluări.

## Predarea sarcinii:

După finalizarea lucrului la sarcină, este necesar să trimiteți mentorului

codul sursă al întregului proiect.

- Denumirea fișierului: `task4_distribution_analysis.py`
- Este obligatoriu să comentați codul, să delimitați clar secțiunile prin întrebări
- Trimiteți un fișier `.zip`, `.rar` sau `.7z` prin intermediul platformei (script + eventual set de date curățat, dacă este disponibil).

## Cerințele de evaluare:

Instructorul va evalua sarcina pe baza următoarelor criterii:

| **Criteriu** | **Procent din notă** |
|---|---|
| Calcularea intervalelor evaluarea televizorului | 20% |
| Calcularea intervalului intercvartil pentru prețurile telefoanelor | 20% |
| Calcularea abaterii standard după brand | 20% |
| Împărțirea în cvartile și analiza vânzărilor pe cvartile | 25% |
| Calitatea și lizibilitatea codului | 20% |

## Mod de evaluare

□□□□□ **(90-100%)**: Sarcina este completă, toate calculele și intrările de date sunt corecte, rezultatele sunt afișate clar, iar codul este lizibil și scris eficient.

□□□□ **(75-89%)**: Erori minore la introducerea datelor sau la calcule, dar majoritatea funcționalității funcționează corect. Prezentarea rezultatelor este clară, cu câteva erori minore.

□□□ **(60-74%)**: Mai multe erori în introducerea datelor, calcule sau afișarea rezultatelor. Codul conține erori minore de logică sau de sintaxă.

□□ **(50-59%)**: Erori majore în introducerea datelor și calcule. Programul funcționează doar parțial, iar afișarea rezultatelor nu este corectă.

□ **(45-49%)**: Sarcina este parțial finalizată, dar există erori semnificative în introducerea datelor și în calcule, care afectează acuratețea rezultatelor. Programul are o funcționalitate limitată și necesită îmbunătățiri pentru a afișa corect informațiile solicitate.

□ **(0-44%)**: Programul nu este funcțional, conține erori grave care interferează cu execuția sa sau sarcina nu a fost înțeleasă corect.

**Succes cu sarcina!**