Task: Analiza de bază a datelor magazinului online

Curs: Data Analysis and Processing using Python
Modul: Analize statistice de bază

Scopul sarcinii:

Scrieți un script Python care utilizează metode de analiză statistică de bază pentru a descoperi tendințe și modele în datele privind cumpărăturile online. Se va practica lucrul cu măsuri de tendință centrală, gruparea și agregarea datelor.

Contextul sarcinii:

Alex vrea să folosească datele pe care le-a colectat din comerțul online pentru a identifica brandurile cele mai vândute, performanța categoriilor și popularitatea produselor. Sarcina voastră este să-l ajutați să obțină informații cheie din date folosind metode statistice de bază și operațiuni agregate.

LINK Academy™

Sarcina:

Vom continua să folosim același set de date pe care Alex a reușit să-l extragă din sistemul informațional al magazinului online. Aceste date pot fi descărcate de aici:

`online_store_data.csv`

Analizați setul de date `online_store_data.csv` și scrieți un script Python care să răspundă la următoarele întrebări:

1. Care este evaluarea medie a produselor din magazinul online?

- Utilizați măsuri adecvate de tendință centrală.

2. Care este cel mai frecvent brand din magazinul online?

- Utilizați măsuri adecvate de tendință centrală.

3. Care este cel mai vândut brand din magazinul online?

- Verificați numărul de produse vândute pentru fiecare dintre branduri. Brandul cu cele mai multe vânzări este cea mai bine vândută.

4. Care este evaluarea medie a produselor pe categorii?

- Presupune folosirea metodei adecvate de calcul a mediei aritmetice, peste coloana cu note, dar pentru fiecare categorie individual.

5. Cum arată popularitatea produselor în funcție de culori?

- Aceasta implică examinarea numărului de unități vândute în funcție de culoarea produsului. Trebuie să afișați numărul total de produse vândute pentru fiecare dintre culori.

6. Care sunt cele mai eficiente 5 branduri din punct de vedere al vânzărilor?

- Trebuie să găsiți cele mai eficiente 5 branduri din punct de vedere al vânzărilor. Eficiența vânzărilor ar trebui măsurată prin raportul dintre numărul de bucăți vândute și numărul total de bucăți achiziționate. Numărul de bucăți achiziționate reprezintă suma unităților vândute și a celor care sunt încă în stoc. Împărțirea numărului de unități vândute la suma dintre unitățile vândute și cele încă în stoc vă oferă eficiența unui brand. Trebuie să folosiți aceste valori pentru a găsi cele mai eficiente 5 branduri.

Ghid pentru rezolvarea sarcinii:

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

1. Care este evaluarea medie a produselor din magazinul online?

- Această sarcină se rezolvă utilizând metoda Pandas pentru calcularea mediei aritmetice pe datele din coloana rating.

2. Care este cel mai frecvent brand din magazinul online?

- Pentru a rezolva această sarcină, este necesar să calculați modul datelor din coloana brand.

3. Care este cel mai vândut brand din magazinul online?

- Rezolvarea acestei sarcini presupune determinarea numărului de unități vândute după numele brandului. Prin urmare, este necesar să grupați datele după coloana brand, apoi să însumați (folosind metoda `sum()`) valorile din coloana `quantity_sold`. La final, trebuie să selectați doar primul element din setul de date obținut. Acesta va fi cel mai vândut brand.

4. Care este evaluarea medie a produselor pe categorii?

- Această sarcină presupune gruparea datelor pe categorii (coloana `category`), apoi calcularea valorii medii pe baza datelor din coloana cu evaluările (`rating`).

5. Cum arată popularitatea produselor în funcție de culori?

- Această sarcină presupune gruparea datelor în funcție de culori (coloana `color`), apoi însumarea valorilor din coloana `quantity_sold`. La final, este necesar să sortați datele în ordine descrescătoare, pentru ca la începutul setului de date să fie culorile cu cele mai multe vânzări.

6. Care sunt cele 5 cele mai eficiente branduri din punct de vedere al vânzărilor?

- Obținerea celor 5 branduri cele mai eficiente din punctul de vedere al vânzărilor presupune gruparea datelor după branduri (coloana `brand`). Pentru fiecare grup, este necesar să executați două operații de agregare:
    - suma valorilor din coloana `quantity_sold`;
    - suma valorilor din coloana `quantity_in_stock`.

Combinarea operațiilor de agregare se poate face cel mai bine folosind funcția `agg()`. După obținerea setului de date, este necesar să creați o nouă coloană. Valorile acesteia vor fi calculate folosind următoarea formulă:

număr_unități_vândute / (număr_unități_vândute + număr_unități_stoc)

Coloana obținută în acest mod va fi utilizată pentru a sorta setul de date, descrescător, pe baza valorilor sale. La final, este necesar să selectați primele 5 rânduri, care vor fi cele 5 branduri cele mai eficiente din punctul de vedere al vânzărilor. Acestea sunt brandurile care au cea mai bună relație între unitățile achiziționate și cele vândute.

Ca ajutor pentru realizarea acestei ultime sarcini, vă poate fi utilă lecția nr. 16, intitulată „Proiect – Analiza datelor cu scopul planificării achizițiilor”. În această lecție a fost realizată o analiză foarte similară pe un set de date dintr-o bibliotecă.

Predarea sarcinii:

După finalizarea lucrului la sarcină, este necesar să trimiteți mentorului codul sursă al întregului proiect.

- Denumirea fișierului: `task3_statistical_analysis.py`
- Structurați codul în blocuri/funcții cu comentarii și nume de secțiuni.
- Trimiteți un fișier `.zip` sau `.rar` prin intermediul platformei (inclus scriptul Python și setul de date dacă sunt curățate).

Cerințele de evaluare:

Mentorul va evalua sarcina pe baza următoarelor criterii:

Criteriu	**Procent din notă**
Utilizarea măsurilor tendinței centrale	20%
Combinarea grupării și agregării	30%
Sortarea datelor	10%
Executarea mai multor funcții de agregare simultan	15%
Crearea unei noi caracteristici	15%
Calitatea și lizibilitatea codului	10%

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!