Tema 1

Analiza exploratorie a datelor pentru oferte de vânzări auto
Mineritul datelor și analiza datelor (MDAD)

2024 - v1.0
Deadline: 02.04.2023 (23:59)

Descriere generală

Pentru această tema va trebui să realizați o analiză exploratorie de date pentru informații legate de oferte de vânzări auto. Rezolvarea corectă și completă a temei presupune implementarea următorilor pași:

1. Citirea și încărcarea datelor din fișierul la dispoziție
2. Transformarea datelor
    a. Din format JSON în formatul necesar pentru restul pipeline-ului
    b. Descoperirea și corectarea erorilor care au apărut din procedura de colectare
    c. Adăugarea sau eliminarea de coloane (acolo unde este cazul, de exemplu prin transformarea celor existente)
3. Analiza datelor obținute
4. Prezentarea analizei realizate într-un raport tehnic

Citirea datelor

Pentru această temă veți avea la dispoziție un set de date cu oferte de vânzare de automobile. Setul de date pe care îl veți utiliza este furnizat sub forma unui fișier text în format JSON. Veți utiliza datele din fișierul respectiv pentru a le încărca, transforma și analiza.

Transformarea datelor

Datele cu care veți interacționa conțin diverse tipuri de erori (pe care va trebui să le descoperiți voi). Unul din obiectivele acestei teme este să descoperiți problemele din setul de date și să le corectați prin metode pe care va trebui să le explicați în raportul tehnic.

Analiza datelor

În analiza datelor sunteți liberi să vă inspirați din ceea ce am discutat la curs și la prezentările practice, însă NU VĂ LIMITAȚI strict la procedurile de acolo. Sunteți

liberi (chiar încurajați) să explorați și alte idei legate de tratamentul și analiza datelor pe care le aveți la dispoziție.

Raport tehnic

Pe lângă fișierele cu codul sursă al temei voastre, va trebui să predați și un raport tehnic care să conțină rezultatele analizei făcute de voi. Raportul tehnic va fi realizat fie sub forma unui fișier de tip PDF conținând toate graficele, tabelele și informațiile din analiza voastră fie sub forma unui jupyter notebook (scris în Python) similar cu cele prezentate la curs. În acest raport tehnic veți descrie, pe lângă rezultatele obținute, procedurile pe care le-ați urmat pentru a colecta, transforma, stoca și analiza datele. Mai mult, împreună cu codul sursă, un alt obiectiv al raportului este să asigure reproductibilitatea experimentelor și tehnicilor de prelucrare și analiză pe care le-ați aplicat datelor. Pe Moodle veți încărca o arhivă *.zip* cu numele vostru și prefixul *Assignment_1_* (e.g. Assignment_1_Ionel_Popescu.zip) în care veți include codul sursă și raportul tehnic.