EXTRAGEREA DATELOR DESPRE PRODUSELE UNUI

MAGAZIN ONLINE FOLOSIND TEHNICA DE WEB SCRAPING

Realizat de: ASAVOAEI ANDREEA-DENISA

Profesor coordonator: ANTON LICZ

UNIVERSITATEA DE VEST TIMIȘOARA,

FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ,

Specializarea INFORMATICĂ ROMÂNĂ, Anul 2, Grupa 1, Subgrupa 1

1. **IDEEA DE BAZĂ**

Programul implementat extrage date cu privire la ofertele de notebook-uri puse la dispoziție pe site-ul web al cunoscutului magazin Altex. Informațiile se referă la titlurile laptopurilor, prețurile actuale, prețurile inițiale – dacă au fost aplicate anumite reduceri cu mențiunile de rigoare –, stocurile produselor, link-urile către produse pentru informații suplimentare și scurte specificații despre acestea. Datele sunt stocate într-un fișier de tip .csv, asemănător cu un tabel, pentru parcurgerea mai facilă și mai rapidă a informațiilor.

1. **DESCRIEREA INSTRUMENTELOR UTILIZATE**

Pentru realizarea acestei aplicații s-a utilizat Python, un limbaj de programare dinamic, care pune accentul pe „curățenia” și simplitatea codului. Acest limbaj dispune de o multitudine de biblioteci și module utile, printre cele utilizate regăsindu-se *urllib.request* și *beautifulsoup4.*

Modulul *urllib.request* definește clase și funcții care ajută la deschiderea URL-urilor într-un univers complex și necesită instalarea pachetului *Requests*.

Biblioteca *beautifulsoup4* vine în ajutorul tehnicii de webscraping cu un parser HTML sau XML care analizează datele primite, de obicei de tip cod HTML. Tehnica de *webscraping* presupune extragerea unor date publice de pe anumite site-uri prin parsarea codului HTML asociat unei pagini. Aceste informații pot fi ulterior prelucrate cu diverse scopuri precum analizarea de date, stocarea acestora într-o bază de date, statistică sau altele.

1. **GRUPUL ȚINTĂ**

Proiectul implementat are drept grup țintă indivizii din orice categorie de vârstă care au acces la un computer cu o conexiune la internet, care au minime cunoștințe de operare a computerului și care sunt interesați fie de ofertele de laptopuri puse la dispoziție de către magazinul online Altex, fie de piața de vânzări actuală.

1. **UTILITATEA ȘI MOTIVAREA**

Prin realizarea acestei aplicații s-a dorit căutarea mai rapidă a produselor disponibile pe site-ul antemenționat prin renunțarea la utilizarea browser-ului care ar putea consuma mai multe resurse decât este de dorit. Se poate economisi o parte a memoriei RAM și chiar timp real deoarece timpul de execuție al programului este de aproximativ 8 secunde, iar deschiderea unui browser și căutarea site-ului, respectiv căutarea categoriei de produse, poate varia în funcție de optimizarea browserului sau calitatea componentelor hardware ale computerului de pe care se operează de la câteva zeci de secunde până la mai mult de un minut.

De asemenea, produsele sunt stocate într-un fisier cu extensia .csv care face parcurgerea mai rapidă pentru un utilizator interesat mai mult de prețuri și specificații decât de imagini ale produselor.

Alegerea acestui magazin online pe baza căruia să fie implementat programul are legătură strict cu felul în care este conceput site-ul, acesta având un cod HTML cu o structură organizată și relativ ușor de parcurs pentru un începător.

1. **STRUCTURA**

**!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!**

**!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!**

**!!!!!!!!!!!!!!!!!!!!!!**

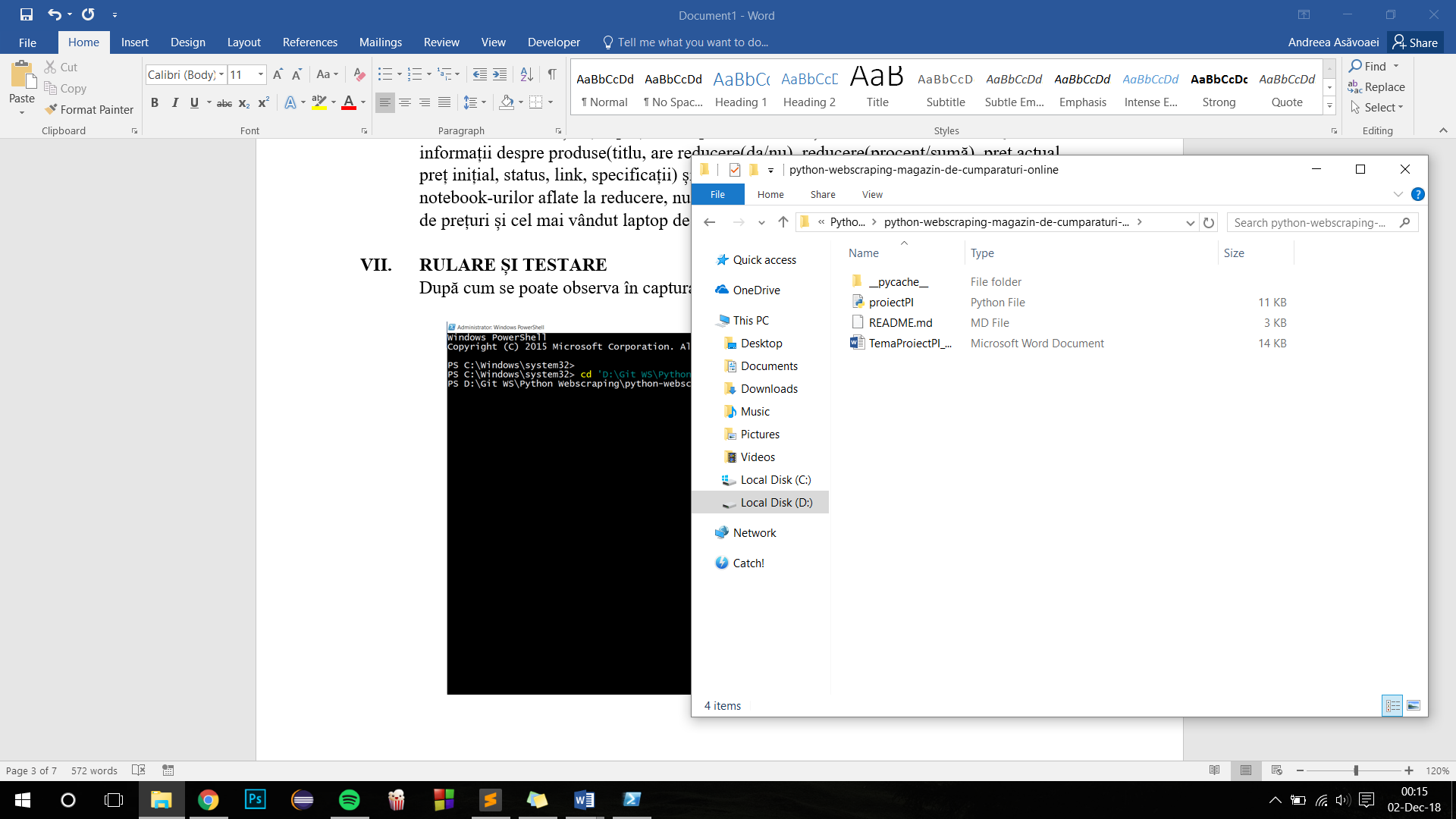
1. **DATE DE INTRARE ȘI IEȘIRE**

Datele de intrare(input) sunt reprezentate de cele 15 pagini cu produse asociate site-ului web ale căror coduri HTML sunt analizate.

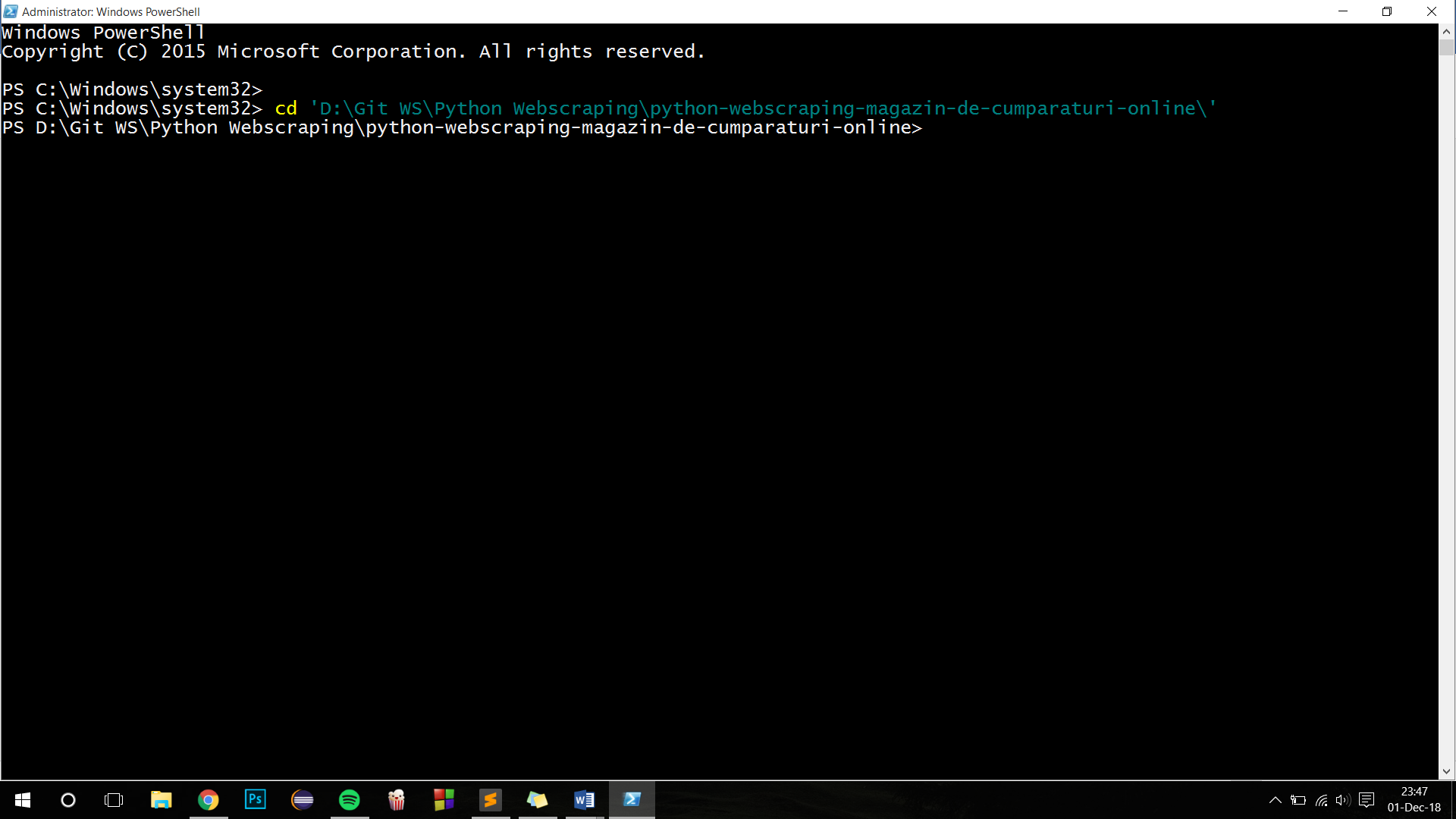
Datele de ieșire(output) sunt reprezentate de fișierul cu extensia .csv conținând informații despre produse(titlu, are reducere(da/nu), reducere(procent/sumă), preț actual, preț inițial, status, link, specificații) și de o scurtă statistică afișată pe consolă cu numărul notebook-urilor aflate la reducere, numărul notebook-urilor situate în anumite intervaluri de prețuri și cel mai vândut laptop de pe site.

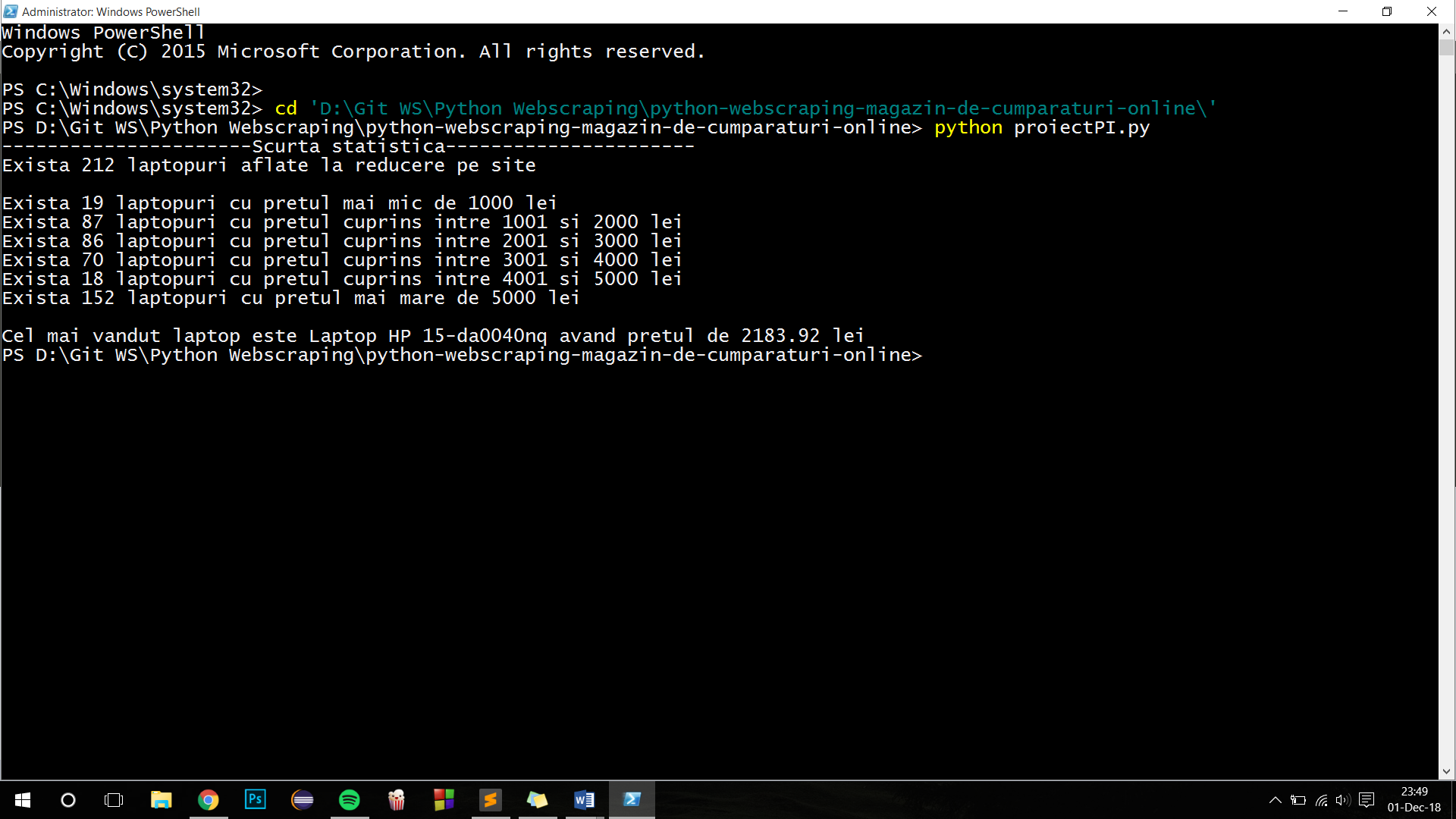
1. **RULARE ȘI TESTARE**

În captura de ecran de mai jos se poate observa conținutul inițial al folderului în care se află programul.

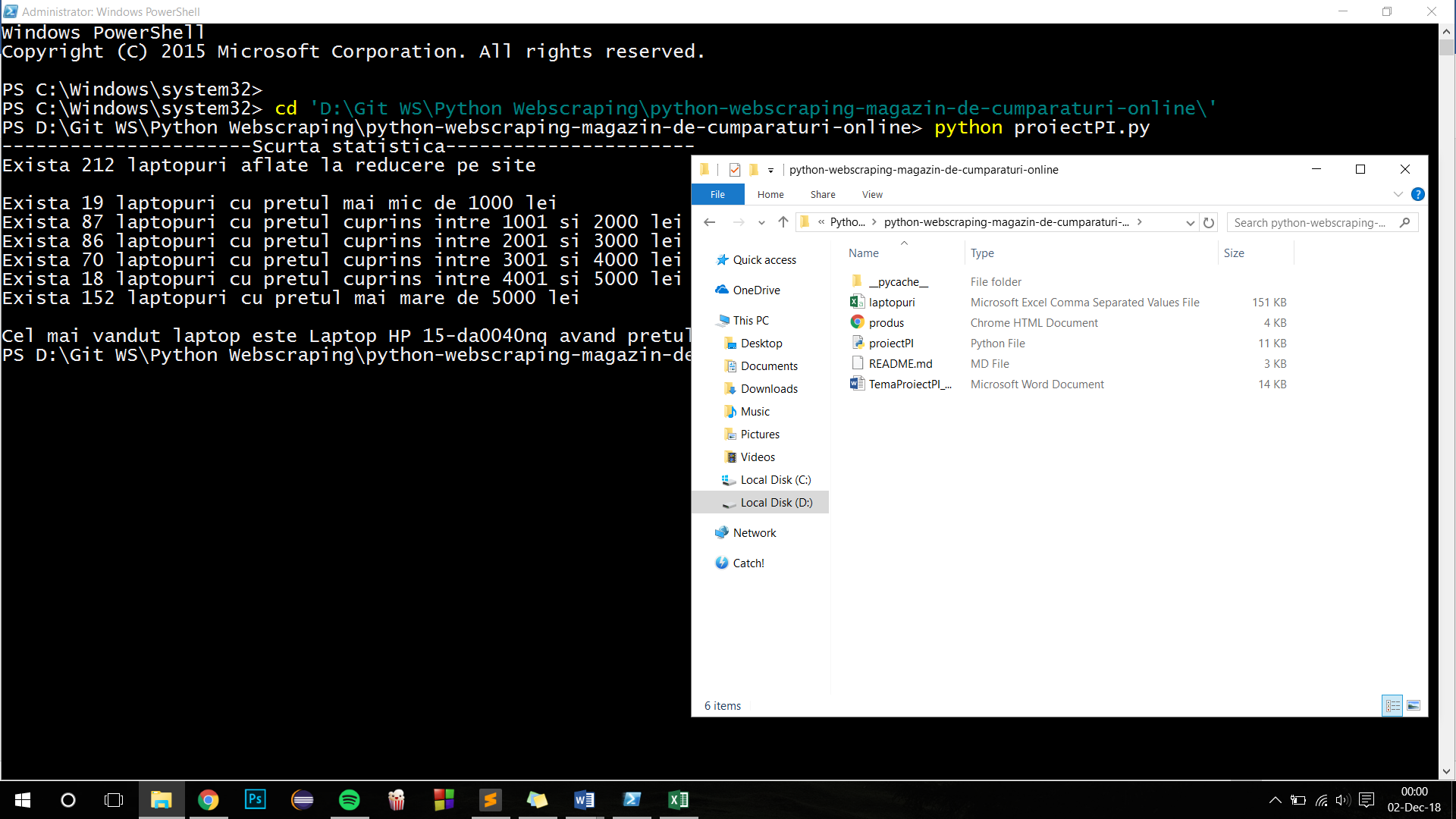


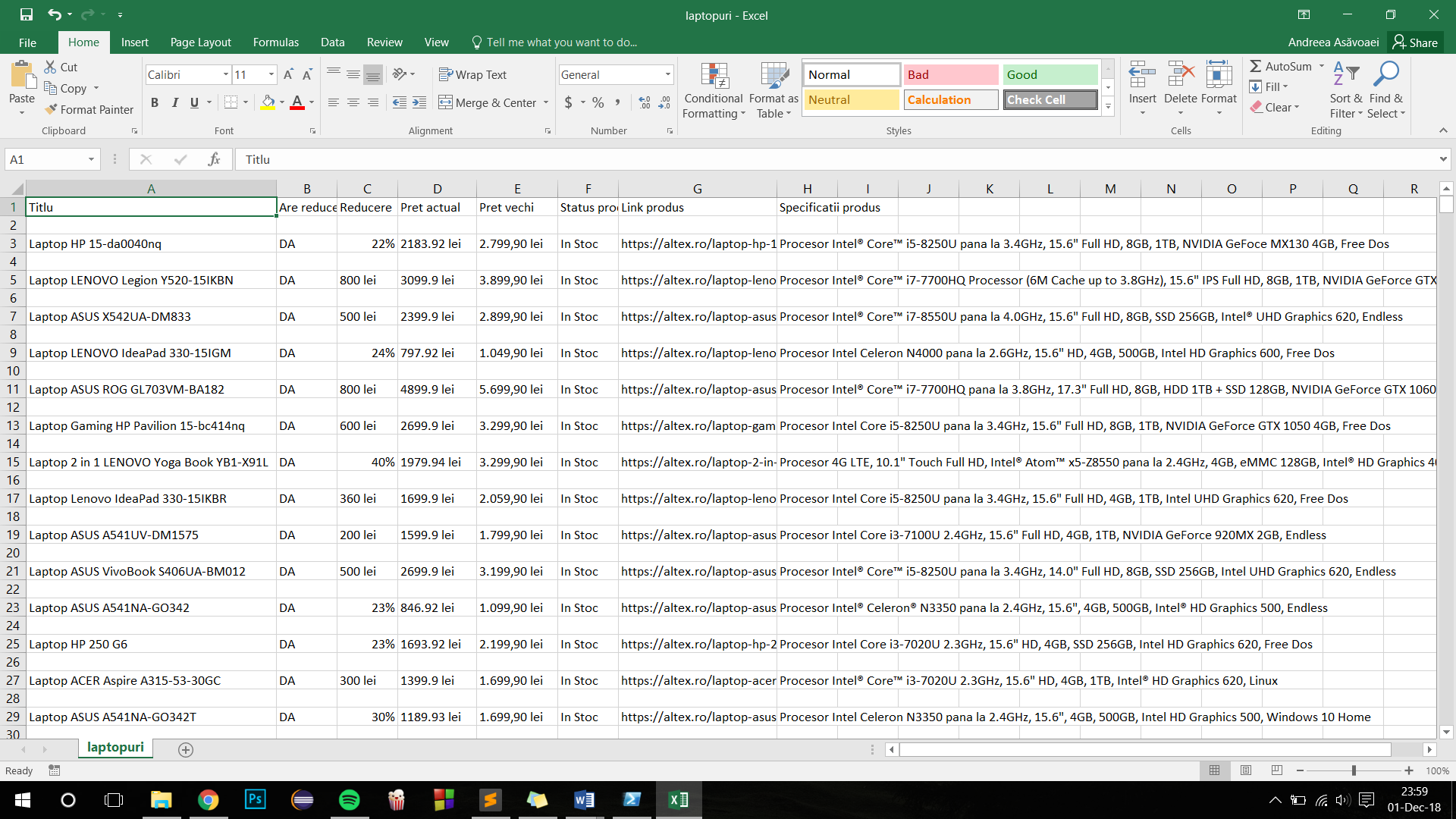
Pentru a rula fișierul *proiectPI.py*, se deschide un terminal în modul administrator. Cu ajutorul comenzii *cd* se accesează folderul în care se află proiectul, iar cu ajutorul comenzii *python* poate fi rulată aplicația, astfel:





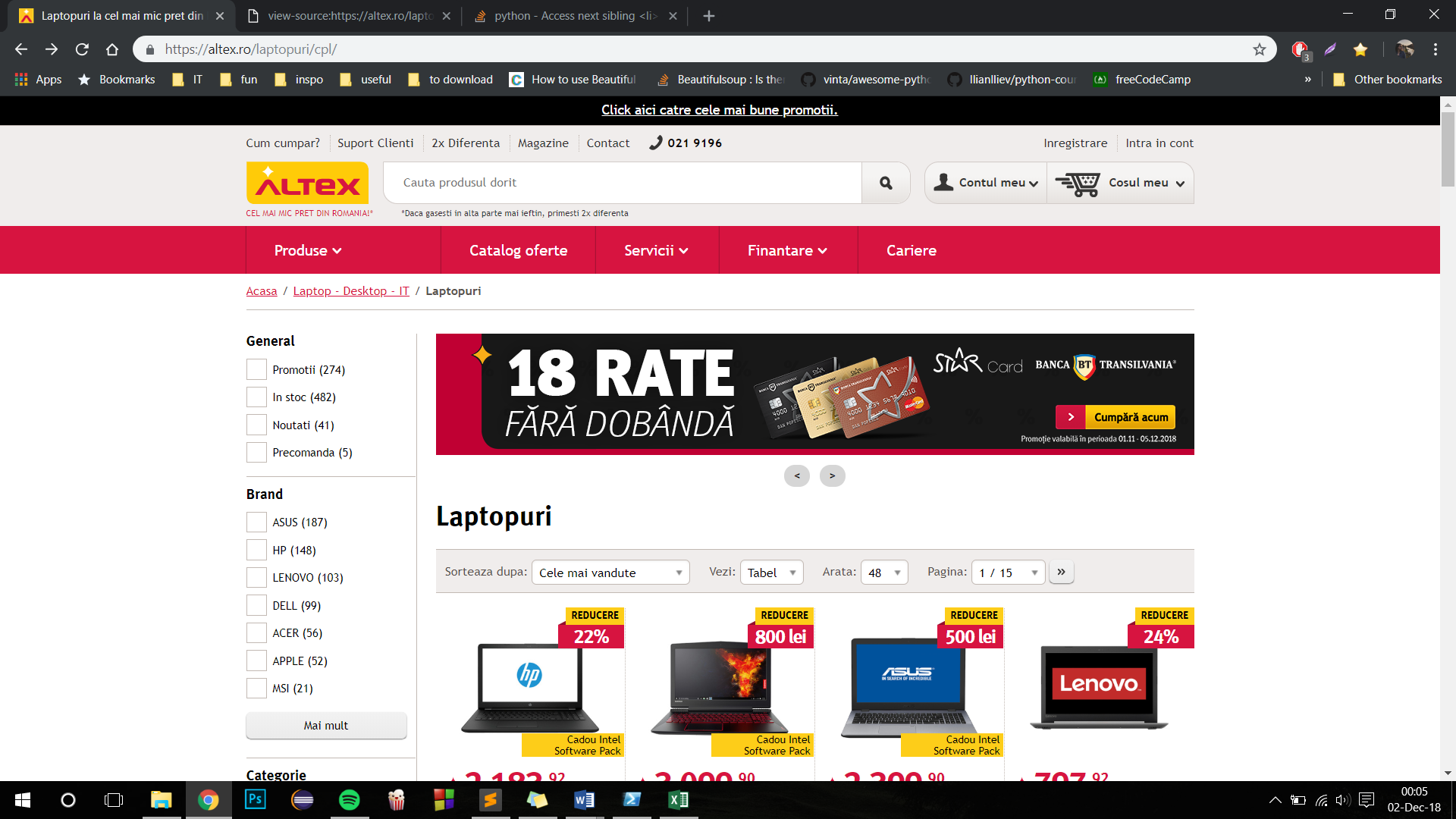
În fereastra de mai sus se poate observa o scurtă statistică a produselor menționată și în secțiunea VI. Se deschide folderul în care se află proiectul și se poate observa apariția unui nou fișier, *laptopuri.csv*, indicat cu săgeată roșie. Acesta poate fi accesat cu ajutorul unui program precum Excel din suita Microsoft Office.

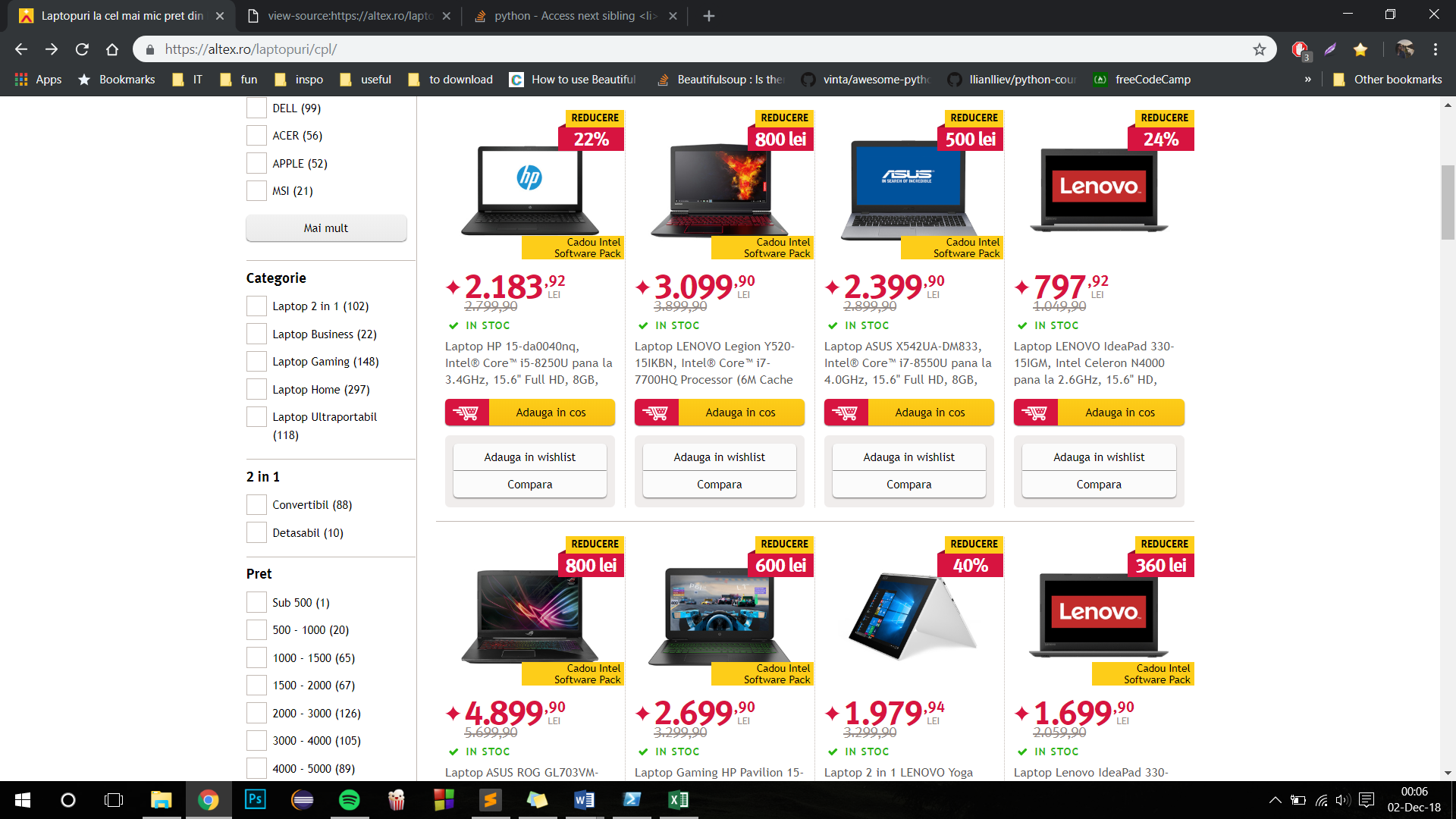


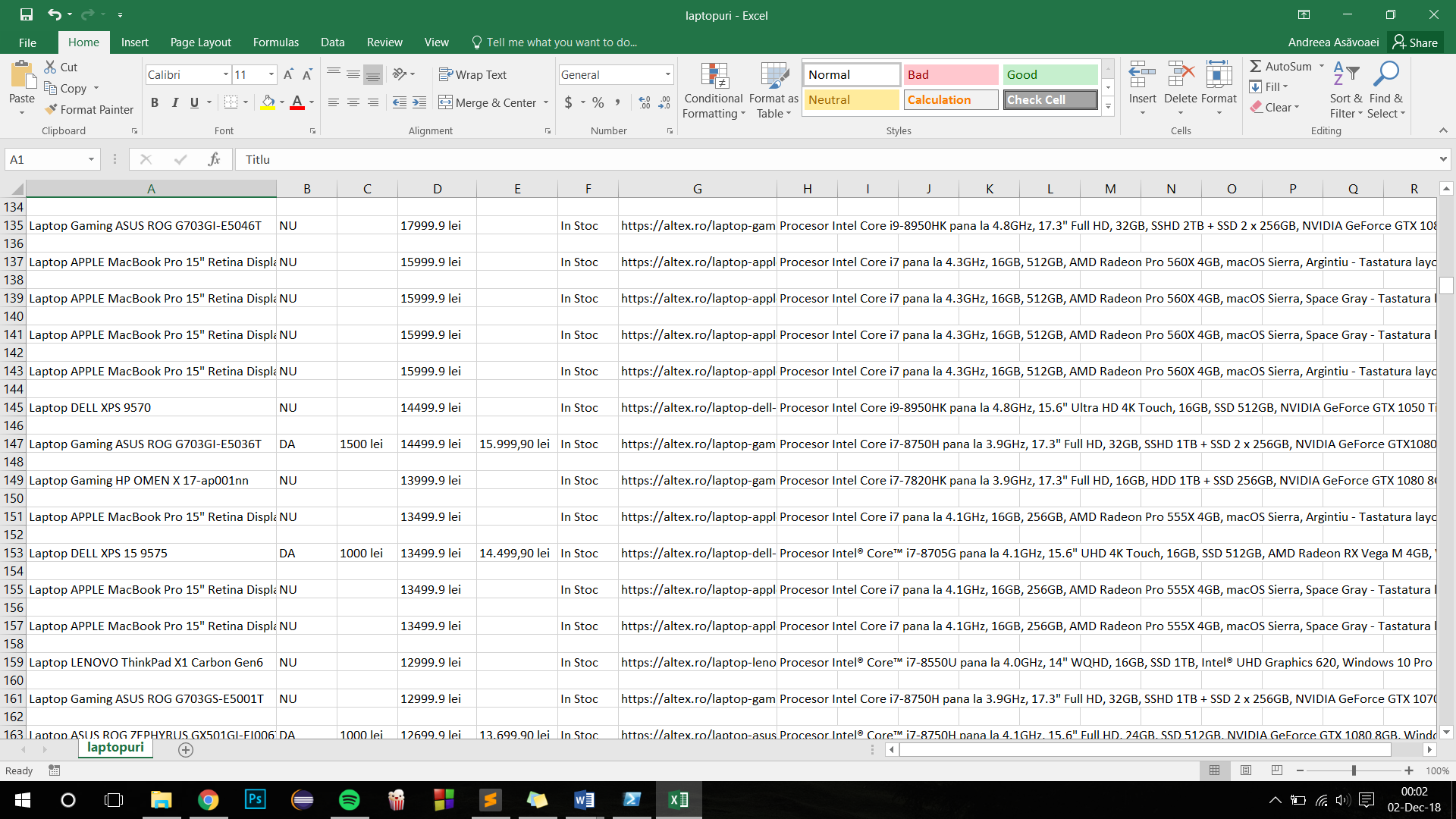


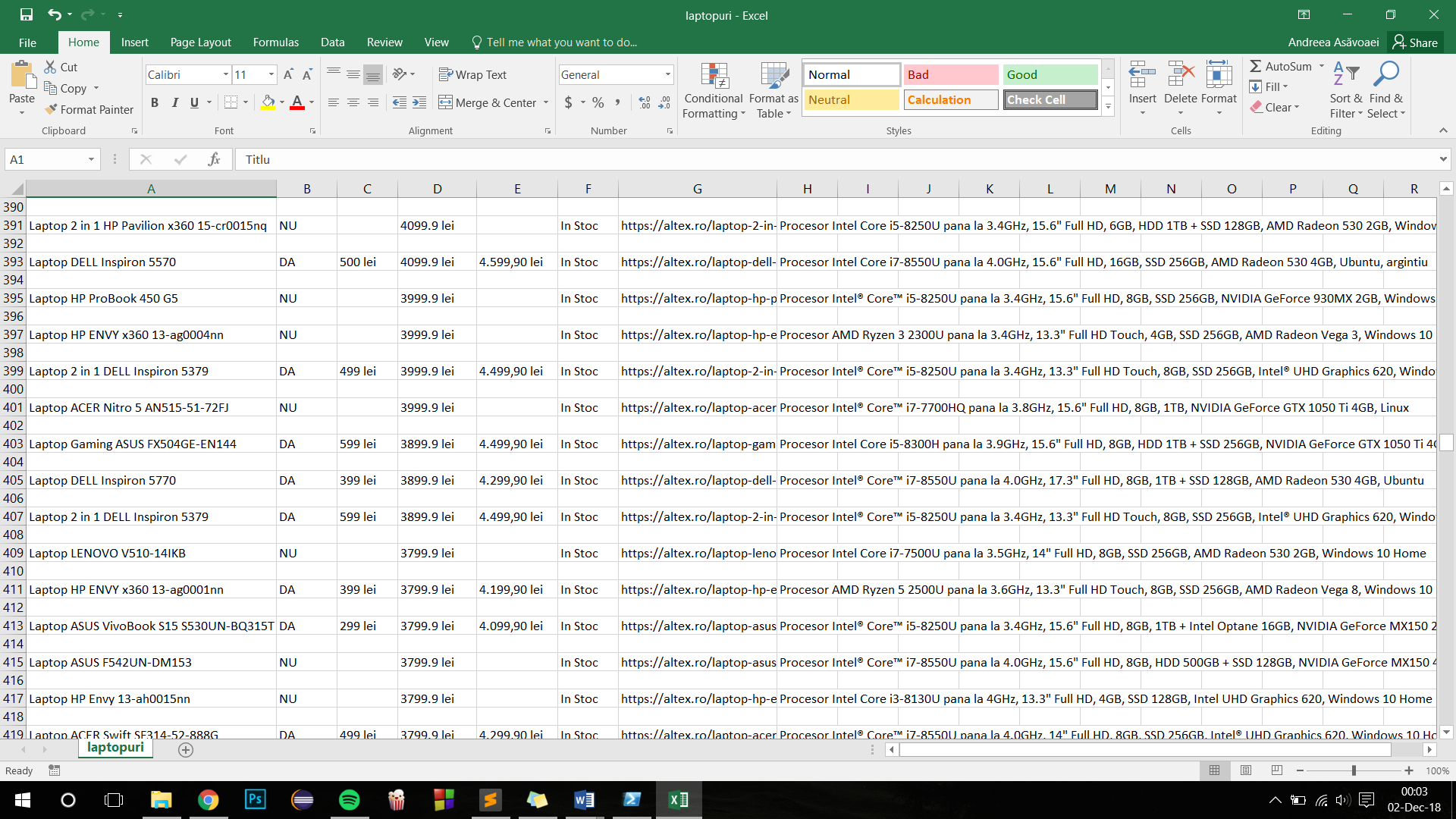
După accesarea fișierului, pentru a compara rezultatele, se poate deschide în browser site-ul magazinului <https://altex.ro/laptopuri/cpl/>, observând că informațiile corespund. Săgețile roșii indică datele care trebuie comparate pentru a sesiza corectitudinea programului. După același șablon pot fi comparate restul produselor listate.

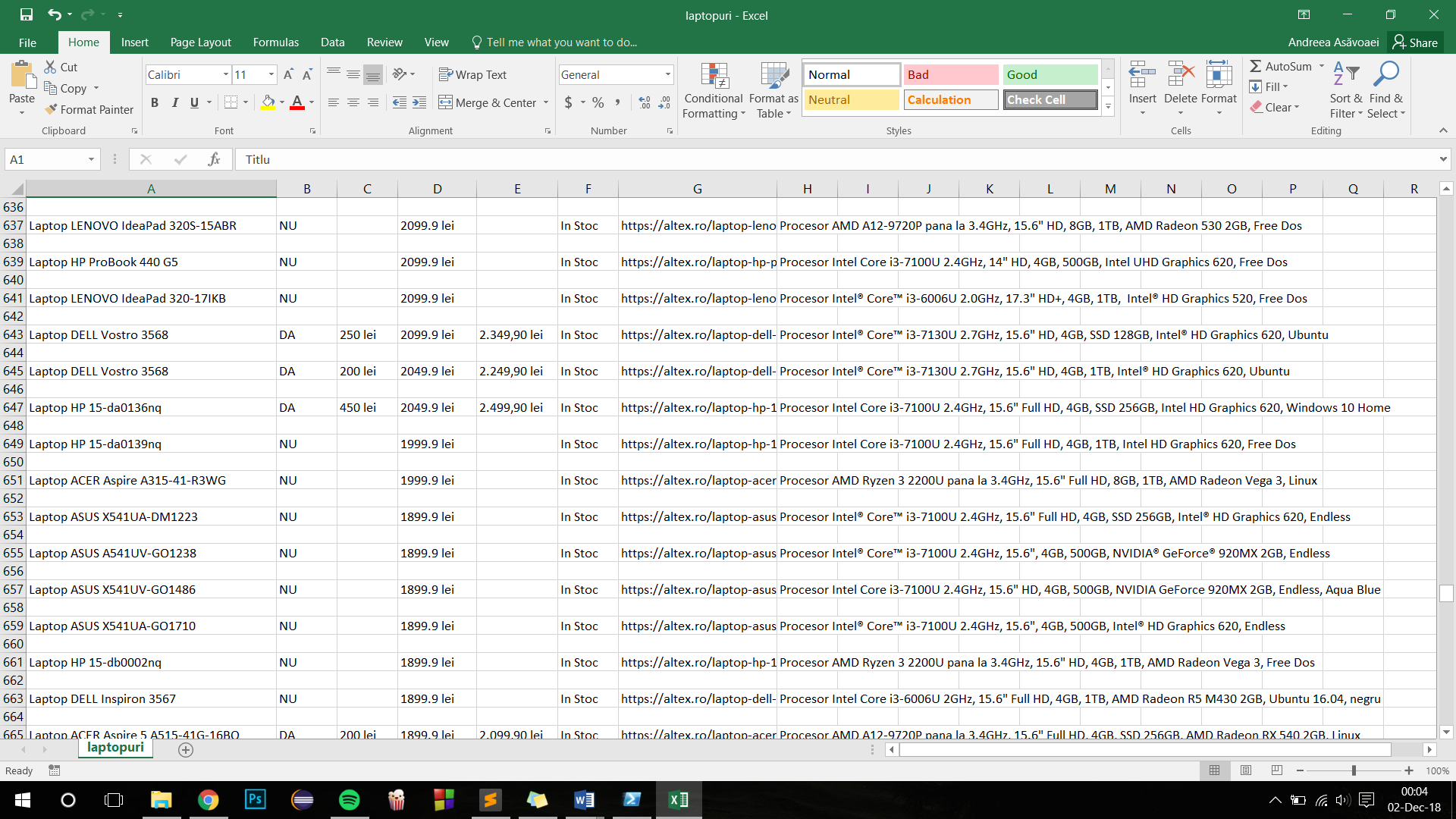
Mențiune: Unele produse din unele pagini ale site-ului web nu pot fi afișate deoarece unor taguri din codul HTML al paginilor respective le lipsesc anumite clase care sunt utilizate la implementarea celorlalte pagini, clase care sunt indispensabile la parsare.











1. **EVENTUALE OPTIMIZĂRI ȘI UPGRADE-URI**

Ca optimizări pot fi incluse: realizarea unei interfețe grafice, posibilitatea de a salva într-un tabel sau într-o bază de date la un anumit interval de timp prețurile produselor pentru a face o statistică cu evoluția acestora, filtrarea notebook-urilor după anumite criterii alese de utilizator, afișarea recenziilor produselor și extinderea aplicației pentru toate categoriile de produse.

1. **CONCLUZII**
2. **BIBLIOGRAFIE**

Site-uri web:

1. view-source:https://altex.ro/laptopuri/cpl/
2. <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
3. <https://docs.python.org/3/library/urllib.request.html>
4. <https://docs.python.org/3/library/csv.html>
5. <https://ro.wikipedia.org/wiki/Python>
6. <https://www.crummy.com/software/BeautifulSoup/bs3/download/2.x/documentation.html>
7. <https://en.wikipedia.org/wiki/Web_scraping>

Cărți:

1. *Web Scraping with Python: Collecting Data from the Modern Web* (Ryan Mitchell, "O'Reilly Media, Inc.", 2015)