Sarcină pentru poziția de inginer de date junior

Avem 3 seturi de date care trebuie unite într-un al patrulea set final. Acesta va conține informațiile relevante obținute din fișierele csv cu informații obținute de pe Facebook, Google sau de pe website-uri. Se pune accesntul pe numele companiilor, adresa acestora, categoria unde își efectuează serviciile și numărul acestora de telefon.

Pentru a rezolva această sarcină, trebuie mai întâi deschise fișierele Excel și observate datele, astfel încât să se știe de la început ce valori există pentru fiecare înregistrare. Deschizând fișierul de pe website, observăm deja o anumită problemă. Toate informațiile necesare sunt puse pe o singură coloană. Până să putem realiza noul set de date, trebuie separate datele în coloane diferite cu ajutorul funcției Text to Columns. Din fericire, acest pas se realizează ușor, deoarece informațiile sunt separate printr-o semicoloană. Această procedură nu trebuie realizată și în celelate două tabele, deoarece în acestea datele sunt deja organizate în coloane.

Pentru a realiza prelucrările necesare, s-a folosit platforma Jupiter Notebook, utilizând limbajul Python 3.12. care este disponibil prin intermediul Anaconda. Seturile de date conțin foarte multe înregistrări și rândurile care nu pot fi citite din cauza formatării greșite sau a incompatibiltății datelor au fost eliminate cu ajutorul funcției disponibile prin biblioteca pandas on_bad_lines. Se va scrie „skip” pentru a trece peste ele.

Mai departe, este important pentru a crea setul de date complet să se observe denumirile coloanelor. Existența mai multor coloane identice va putea simplifica complexitatea tabelului care va fi creat. Văzând denumirea coloanelor în cele trei fișiere, se poate constata că există mai multe atribute comune precum domeniul, țara, regiunea, numărul telefonic, orașul și categoria. Ele însă sunt notate diferite în tabele, așadar aceste coloane trebuie mai întâi redenumite ca să avem punctul comun de plecare.

După ce redenumirea a fost realizată cu succes, se poate trece la îmbinarea seturilor de date într-unul singur. Se va face o joncțiune exterioară (reuniune), deoarece se dorește a se păstra fiecare informație utilă pe care putem să o aflăm despre companiile existente. Se vor și rearanja coloanele, astfel încât mai întâi să apară valorile comune dintre cele trei tabele, apoi cele unice sau care apar în dooar două din ele. Jonțiunea se va realiza pe baza atributului domeniu, acesta fiind foarte important prin faptul că oferă accest direct către pagina web a companiei. După acesta se vor afla următoarele atribute în această ordine: nume, categorie, telefon, oraș, țară, regiune, codul țării. La final, vor urma și atributele care nu sunt comune în toate cele trei seturi de date.

După ce se obține setul cel mare de date, acesta trebuie prelucrat mai departe, fiindcă el conține multa informatie redundantă. Pentru a completa în mod corect coloanele principale, s-au foarmat coloane intermediare de unde puteau fi preluate datele. Acestea sunt primele care vor fi

șterse, deoarece rolul lor a fost doar temporar, ele fiind necesare pentru a fi portate in mod corect in setul mare de date. Cu un set de date mai aerisit, se observă că nu au fost portate toate informațiille în tabelul cel mare, astfel se vor introduce și valorile pentru elementele care aparțin doar de un tabel sau de două maxim. Odată ce acest pas a fost completat, se vor identifica atributele care nu au nicio relevanță pentru o analiză de date sau pentru informarea completă unei persoane.

După ce în set au rămas doar valorile importante, se va trece la analiza mai aprofundată a acestuia. Se vor căuta posibile înregistrăti care sunt duplicate sau foarte similare. Ele reprezintă un conflict de date foarte important, fiindcă pot apărea confuzii legate de acuratețea informațiilor. Dacă acestea există, mereu se va căuta înregistrarea care este cea mai completă și care are cele mai multe date. Restul rândurilor similare sau duplicate se vor șterge. De asemenea, se vor identifica și valorile lipsă în coloanele comune. Ele vor crea alt conflict de date, deoarece lipsesc informații vitale despre o anumită firmă. Aceste spații goale se vor completa cu ajutorul funcției fillna, fie cu stringul „ffill”, fie cu șirul „bfill”, acestea fiind completări automate pentru valorile categorice. Doar acestea se vor aplica, deoarece în setul de date nu există valori numerice.

Curățarea și corectarea setului s-a bazat pe identificarea înregistrărilor problematice și eliminarea rândurilor care nu oferea suficiente informații. Înregistrările rămase sunt cele mai complete și oferă cele mai multe informații relevante pentru o viitoare analiză. Este foarte important ca în cadrul unui set se date să se lucreze cu înregistrări care oferă cât mai multe informații relevante, fiind astfel mult mai ușor să se creeze predicții pentru care se pot crea planuri. De asemenea, se pot crea mai repede și mai la îndemână rapoarte sau analize pentru o anumită ramură a sectorului economic sau pentru a avansa în cadrul unei investigații.