Data Opsiku Mariolko, nasza firma się pali, nikt nie wie co robić, potrzebujemy twojej pomocy!

Nasz klient – Szampańskie Zdrowie Nikt Się Nie Dowie Since 2007 – zaczął przysyłać nam swoje tygodniowe dane w formie plików .csv spakowanych w jeden .zip.

Pliki mogą mieć różne separatory, informacja o separatorze jest zawarta w pliku metadata.txt w formie „nazwa pliku: <separator>”. Jeśli w pliku metadanowym nie pojawiła się nazwa pliku, to znaczy że domyślnie separatorem jest przecinek.

Uwaga, w plikach zip oprócz naszych plików csv oraz pliku metadanowego mogą być jakieś inne pliki, np. jsony czy inne txt. Musimy zignorować je, one nas nie obchodzą.

Niestety nie wiemy jakie są nazwy kolumn w plikach csv – zawsze są one inne poza kilkoma wyjątkami (np. active, event\_datetime, sales czy brand)

To nie koniec tarapatów, nastąpiła awaria w systemie generującym dane klienta, przez następne 3 lata pliki z danymi mogą być wybrakowane, tj. może brakować w nich kolumn w niektórych rekordach. W takiej sytuacji powinniśmy pomijać takie rekordy, ale równocześnie powinniśmy monitorować liczbę błędów, jaka wystąpiła w każdym pliku. Na końcu musimy zapisać do osobnego pliku taki raport w którym wypiszemy nazwę każdego pliku oraz liczbę napotkanych błędów. Dodatkowo, jeśli w pliku więcej niż 30% rekordów będzie błędnych, to musimy dopisać CRITICAL w logu o błędach.

Poprawne rekordy w plikach chcemy przetworzyć na dataframe’y pandasowe.

Tak powstałe dfy musimy odfiltrować – chcemy tylko rekordy aktywne (spójrzmy na kolumnę active) i tylko rekordy z ostatnich 2 lat (event\_datetime; powinno się to przesuwać – zawsze chodzi nam o 2 lata względem dzisiejszego dnia).

Po takim odfiltrowaniu połącz powstałe dataframe’y – wybierz tylko kolumny brand i sales i zrób jeden dataframe gdzie w kolumnach brand i sales są wyniki z wszystkich poprzednich dataframe’ów jedne pod drugimi. Następnie zrób sumę pogrupowaną według brandów.

To będzie nasz dataframe końcowy. Pozostaje tylko jedna rzecz – przeiteruj po wierszach i zapisz do pliku csv o nazwie sales.csv brandy oraz ich salesy.

Wszystko powinno być w formie uniwersalnych funkcji. Na końcu powinna być funkcja o nazwie „run”, która przyjmuje nazwę pliku zip, a także nazwę folderu docelowego, w którym będzie zapisany output, i przeprowadza całą analizę.