Aleksandra Jarzyńska 136722 (praca samodzielna – 5h + 4h + 3h)

1. Opis koncepcji rozwiązania problemu.

Wykorzystany Python i biblioteka pandas – rozwiązanie oparte o DataFrames.

W pierwszej kolejności wczytuję pliki i zapisuję je jako DataFrame.

Wykonuję .uppercase() na polach tekstowych, co pozwoli na wyeliminowanie błędów przy mergowaniu wynikających z różnej wielkości znaków w danych.

W przypadku plików „a” i „c” podczas odczytywania korzystam z podanej informacji o podanej długości poszczególnych pól i parsuję stringa – jednocześnie pozbywam się białych znaków na końcach, w przypadku plików „b” korzystam z wbudowanej funkcji .read_csv(), która pozwala na określenie rozdzielnika pól.

Na podstawie pliku „c” wyznaczam klientów, których dochód przewyższa przekazany w parametrze INCOME_THRESHOLD, a następnie szukam wśród klientów z najwyższym dochodem, tych których dochód przekracza VIP_INCOME – usuwam ich 200 (ponieważ w danych jest ich więcej niż 200).

Dla plików „a” – wyznaczam sumę transakcji, grupuję po ID klienta i tworzę dwie dodatkowe kolumny – jedną z sumą wartości zakupów, drugą z sumą wartości zwrotów – na koniec odejmuję zwroty i zapisuję rezultat w kolumnie „sumA”.

Podobnie z plikami B – rezultat zapisuję w „sumB”.

Merguję (outer join) DataFrame dot plików A i B po imieniu, nazwisku i adresie, następnie wyliczam wspólną sumę dla transakcji z plików A i B – zapisuję jako „PURCHASES”.

Wynik merguję (outer join) z wynikiem z pliku „c” - to daje mi ostateczny rezultat, gdzie wprowadzam jedynie drobne modyfikacje – usuwam niepotrzebne kolumny, ustawiam wartości nieokreślone zamiast NaN – taki dataframe zapisuję do pliku „out.txt” rozdzielając pola znakiem „|”.