Уважаемый Кандидат!

Мы рады приветствовать Вас на этапе тестового задания на позицию Data Scientist. Мы надеемся, что наше задание будет для Вас интересным, и в ходе его решения Вы сможете проявить имеющиеся у Вас компетенции в области анализа данных.

Вам предоставляется выгрузка регистраций 2298 абонентов мобильной связи на базовых станциях в городе Москва в течение одного месяца. Формат данных следующий:

LAC	CID	TS	FULLDATE	HASH_ID
Код зоны мобильного оператора	Идентификатор соты	UNIX-время - способ кодирования времени, принятый в UNIX операционных системах. Означает число секунд, прошедших с 01.01.1970 до рассматриваемого события	Дата	Идентификатор абонента мобильной связи

Выборка логически разделена на две части:
1) с 1 по 14 число месяца
2) 16 по 30 число месяца

В каждой части используется своя система присвоения идентификаторов абонентам, что приводит к тому, что один и тот же абонент в разных частях выборки имеет разные идентификаторы: с 1 по 14 число – один, с 16 по 30 число – другой.

Ваша задача:
*разработать алгоритм, который будет устанавливать уникальное соответствие между идентификаторами из первой и второй части выборки. Ожидаемый результат – таблица соответствий вида id1 – id2 в формате *.csv, где id1 – идентификатор из первой части выборки, id2 – идентификатор из второй части выборки. В помощь Вам предоставляется эталон из 491 пары id1 и id2. Вам необходимо определить оставшиеся 1000+ соответствий.*
Результат просим оформить в следующем виде:
а) алгоритм в формате Python-ноутбука или R-скрипта
*б) презентация в формате *.ppt или *.pdf, в которой кратко изложить суть алгоритма*
*в) результат работы алгоритма: таблица соответствий в формате *.csv.*

Данные Вы сможете скачать по ссылке:
https://www.dropbox.com/s/9v3l2olj621djph/%201F_test.zip?dl=0

Желаем успешного выполнения задания!

ООО "ЕдиныйФактор"
Москва, Деловой Центр «Империя»
Пресненская наб., д. 6, стр. 2
+7 (499) 643-83-25
reception@onefactor.com
onefactor.com