Компания Киевстар предоставляет ряд дата-продуктов для внешних клиентов, основанных на анализе анонимизированных данных о мобильности (перемещениях) абонентов. Для этого важно понимать, какие существуют группы абонентов, исходя из их профиля перемещений за определенный период времени.

В рамках курсового проекта каждой группе нужно показать результат по 2-м поставленным задачам.

ЗАДАЧА №1.

На основании набора данных по одним и тем же абонентам за период 04.2019-09.2019 необходимо провести анализ и применить алгоритмы машинного обучения без учителя с целью выделения хорошо отличающихся сегментов абонентов.

Также у вас есть данные по этим же абонентам и за "карантинный" период 04.2020-09.2020, чтобы вы могли проанализировать изменения в мобильности абонентов, вызванные ограничениями, введенными государством для противодействия распространению эпидемии.

В итоге мы ожидаем получить сравнение показателей мобильности за периоды 2019 и 2020 для каждого из полученных сегментов.

ИСТОЧНИКИ ДАННЫХ:

- **Датасет перемещений между районами** (user_routes_2019_2020.csv) содержит данные о перемещениях абонентов между районами в 2019 и 2020 годах. Под перемещением подразумевается, что абонент до начала поездки и по окончанию поездки находился в соотв. районе более 3-х часов.
  Отфильтрованы абоненты, которые не перемещались за пределы своего района или перемещались, но на время менее чем 3 часа.
  Дополнительное условие - все абоненты имеют хотя бы 1 перемещение в 3 месяцах из 6.

Структура:

user_id - уникальный идентификатор абонента

start_area_id - ключ района старта поездки

finish_area_id - ключ района окончания поездки

start_time - время старта поездки в формате yyyy-MM-dd HH: 00:00

finish_time - время окончания поездки в формате yyyy-MM-dd HH: 00:00

hmonth - месяц начала поездки в формате yyyy-MM-dd

- **Справочник районов Украины** (districts_info.csv) содержит справочную информацию о районах - название, координаты, список "соседних".

Структура:

area_id - ключ района

region_name - название области

area_name - название района

centroid_lat - широта центроида района (крупнейшего города в районе)

centroid_lon - долгота центроида района (крупнейшего города в районе)

neighbor_area_idx - ключи "соседних" районов, имеющих общие границы с текущим

Ожидаемый результат по задаче:

1. разработка сегментации на данных 2019 года и ее презентация;
2. описание переменных, используемых в сегментации и профилирование сегментов по этим переменным;
3. расчет аналогичных переменных на данных 2020 года;
4. анализ динамики изменения для каждого сегмента и визуализация результатов.

ЗАДАЧА №2.

У вас есть данные всех транзакций абонентов за 1 месяц, которые фиксировались в одном из городов (у каждой команды свой город), а также данные по перемещениям этих же абонентов за 2019 год аналогично задаче №1. Необходимо выделить аналогичные сегменты для этой выборки абонентов и для каждого сегмента, на основании транзакционных данных, проанализировать каким образом они въезжали и выезжали из города.

ИСТОЧНИКИ ДАННЫХ:

- **Датасет с транзакциями** (user_transactions_<city_name_eng>.csv) содержит данные по всем транзакциям абонентов за календарный месяц. При этом для каждого города отбирались абоненты, которые в нем не проживают, но в течение месяца имели хотя бы 1 транзакцию в городе.

Структура:

user_id - уникальный идентификатор абонента

event_dt - время совершения транзакции в формате yyyy-MM-dd HH:mm:ss

lat – широта

lon - долгота

- **Датасет перемещений между районами** (user_routes_<city_name_eng>.csv) содержит данные о перемещениях абонентов (из датасета с транзакциями, ключи совпадают) между районами в 2019 году. Под перемещением подразумевается, что абонент до начала поездки и по окончанию поездки находился в соотв. районе более 3-х часов. Отфильтрованы абоненты, которые не перемещались за пределы своего района или перемещались, но на время менее чем 3 часа.
  Дополнительное условие из задачи №1 (все абоненты имеют хотя бы 1 перемещение в 3 месяцах из 6) - не применялось.

Структура:

user_id - уникальный идентификатор абонента

start_area_id - ключ района старта поездки

finish_area_id - ключ района окончания поездки

start_time - время старта поездки в формате yyyy-MM-dd HH: 00:00

finish_time - время окончания поездки в формате yyyy-MM-dd HH: 00:00

hmonth - месяц начала поездки в формате yyyy-MM-dd

- **Полигон города** (<city_name_eng>.geojson) содержит данные о "границах" города в формате geojson (https://en.wikipedia.org/wiki/GeoJSON).

Ожидаемый результат по задаче:

1. применить сегментацию, разработанную в задаче №1 к абонентам задачи №2 и проанализировать структуру полученных сегментов, сравнить с полученной в задаче №1, визуализировать результаты;
2. в разрезе сегментов выделить ключевые точки въезда/выезда из города по транзакционным данным, для каждой из точек собрать статистику количества абонентов и визуализировать результаты;
3. на транзакционных данных разработать сегментацию по целям посещения города и визуализировать результаты.