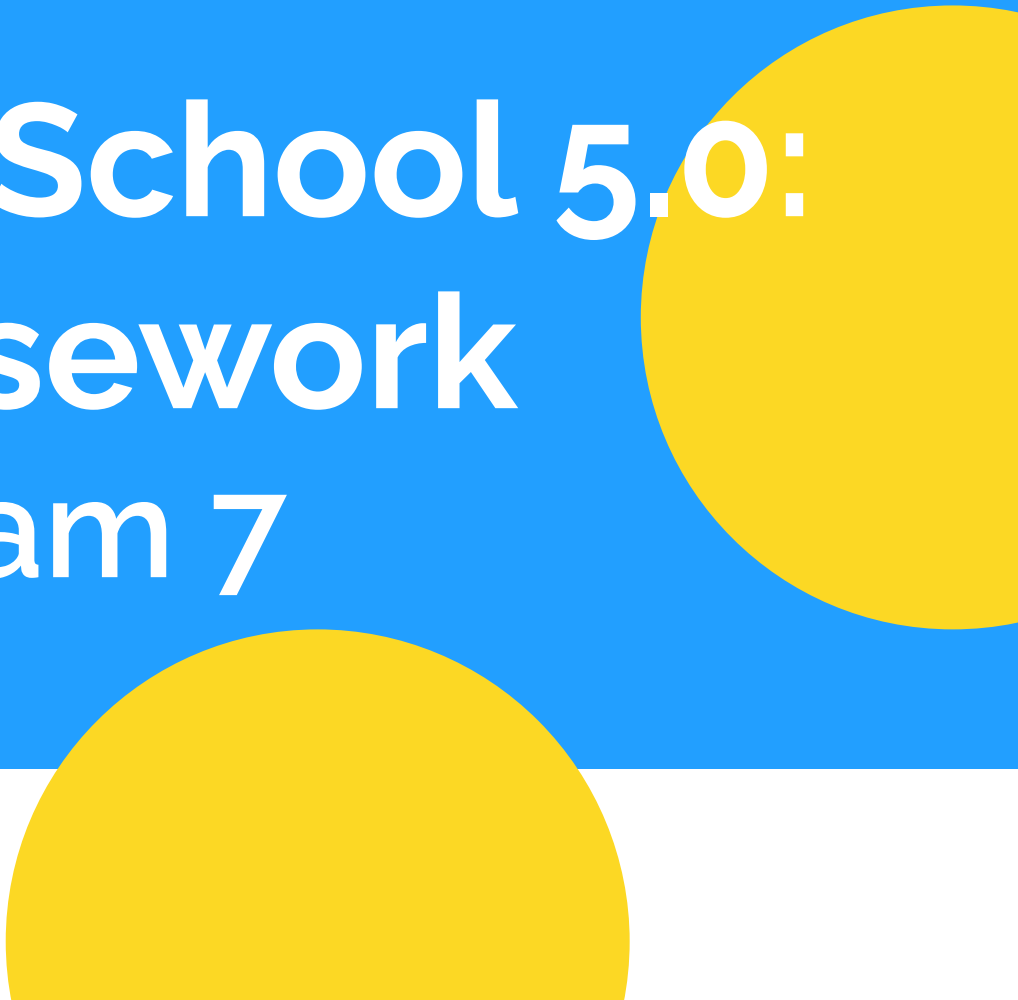


Big Data School 5.0:

Coursework

team 7

The background of the slide is a solid blue color. On the right side, there is a large yellow circle that is partially cut off by the edge of the frame. At the bottom center, there is another large yellow circle, also partially cut off by the bottom edge of the frame.

Команда ●●

- **Самошин Андрій** - team leader, machine learning, analytics, presentation
- **Пилипович Максим** - machine learning, analytics
- **Мельник Ольга** - machine learning, analytics, statistics
- **Черниш Лідія** - analytics, presentation

Завдання 1: вхідні дані ●●

- **Дано:** Датасет переміщень за 2019 та 2020 рік (період 4-9 місяці)
- **Необхідно:** Проаналізувати зміни показників мобільності
- **Доповнення:** вирішено порівнювати не повні періоди, а окремо під час та після локдауну 2020 року



Підготовка даних ●●

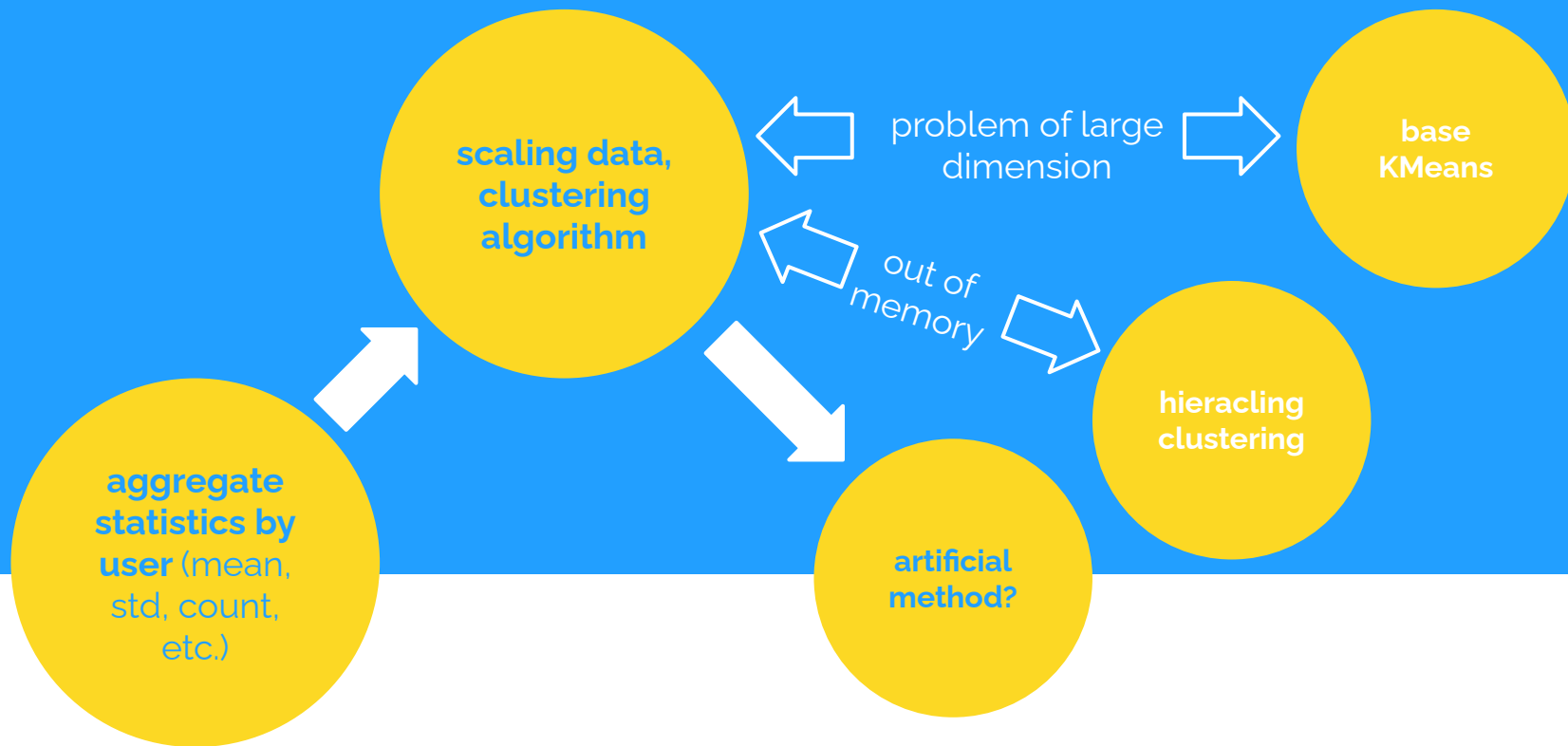
Базові ознаки:

- time based features
- join districts info

Розширені ознаки (на основі гіпотез):

- Переміщення в межах області
- Ймовірнісне визначення домашнього району
- Повний (частковий) робочий день користувача (**умови:** будній день, дельта між поїздками 3-6 годин, початок наступної поїздки не раніше 18:00)

Підходи до вирішення ●●



Рішення завжди є ●●

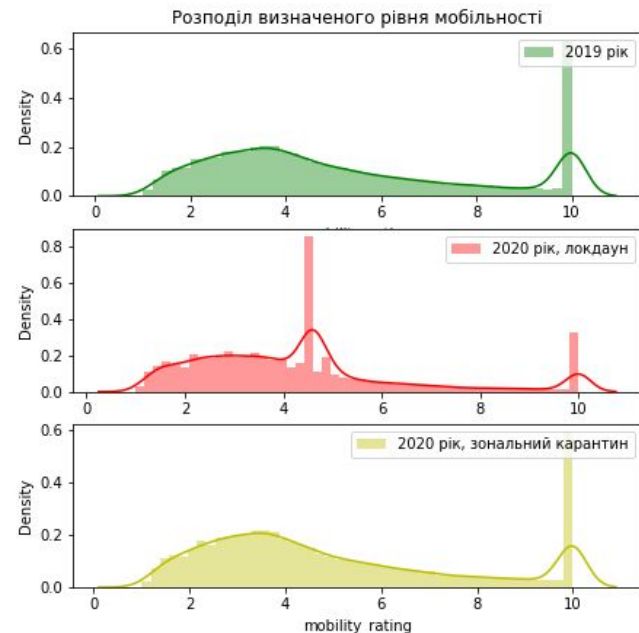
Ідея: Вручну зменшити розмірність шляхом введення агрегуючого індексу мобільності

- **mobility_index** = $1 + 9 * (\text{mobility_score in range } [0,1])$

$\text{mobility_score} = (\text{MD} * \text{SR} + \text{MND} * (1 - \text{SR})) * ((1 - \text{FT}) + (1 - \text{PT}))$

- прогнозована зміна розподілу під час локдауну (-32% частки індексів 5+), але повернення після

KMeans для mobility_index та route_count

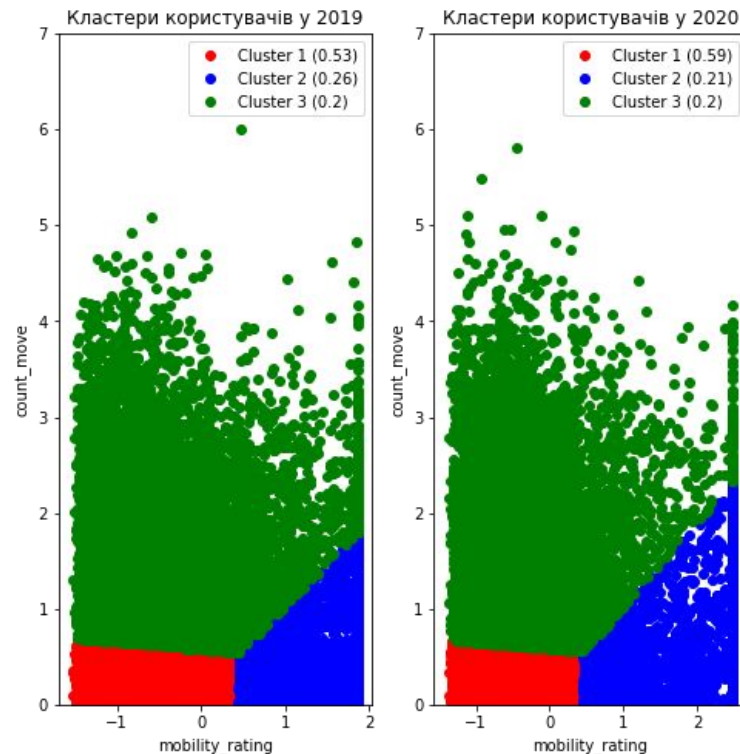


Результати сегментації ●●

- Обрана кількість кластерів - **3**
- **Silhouette score**: 0.62

Отримані співвідношення
(2019/2020 рік):

- сегмент 1 - **53% / 59%**
- сегмент 2 - **26% / 21%**
- сегмент 3 - **21% / 20%**



Характеристика сегментів ●●

Сегмент	Профіль
Кластер 0	<ol style="list-style-type: none">1. Середня мобільність, але мають багато екстремальних значень2. Невелика кількість коротких переміщень3. (Гіпотеза) Більше 60% - робочі поїздки4. 80% поїздок додому із сусідніх районів
Кластер 1	<ol style="list-style-type: none">1. Дуже висока мобільність (8.7)2. Не так часто переміщуються (через довгі переїзди)3. (Гіпотеза) ~5% поїздок на роботу повного дня4. Майже в 3 рази триваліші поїздки (120+ км)
Кластер 2	<ol style="list-style-type: none">1. Середня мобільність (3.5)2. Велика кількість переміщень (200+) з підвищеною дисперсією3. Нейтральні поїздки недалеко - 30-40 км4. Понад 80% поїздок додому - після 18:00 по будням

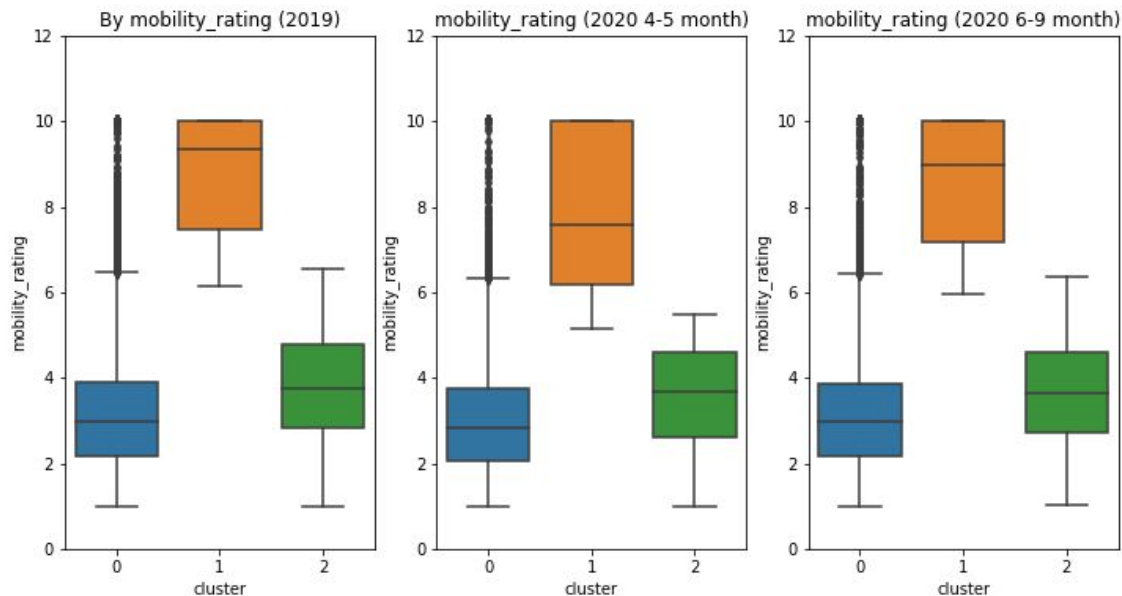
Опис порівняльних ознак ●●

Обраний індикатор	Характеристика
Count of moves	Сума всіх переміщень користувача за відведений час. Агрегується до дня. Основне використання: допомогти в потенційному масштабуванні інших індикаторів в разі раптових змін в шаблонах поведінки
Mean distance traveled (поза домашнім районом)	Середня відстань, подолана кожним користувачем в інші райони. Переваги: крім загального переміщення, відображає рухи поза домашньою адміністративною зоною. Недоліки: низька точність тому, що вважаємо відстань між координатними центрами, а не фактичними координатами переміщення
Mobility rating	Акумуляована оцінка активності користувача. Переваги: відображає активність по всім фронтам поїздок, які не вважаються робочими

Порівняння та динаміка ●●

Помітні зміни для 1 кластеру: майже **на 1.5 позицій зменшилась медіана**.

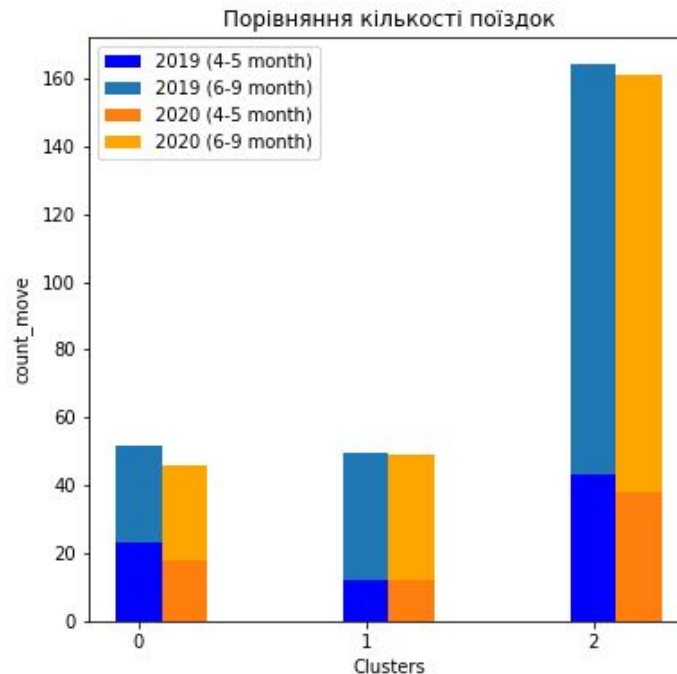
Інші, і без того малоактивні кластери, **зменшили дисперсію рейтингу на 10%**



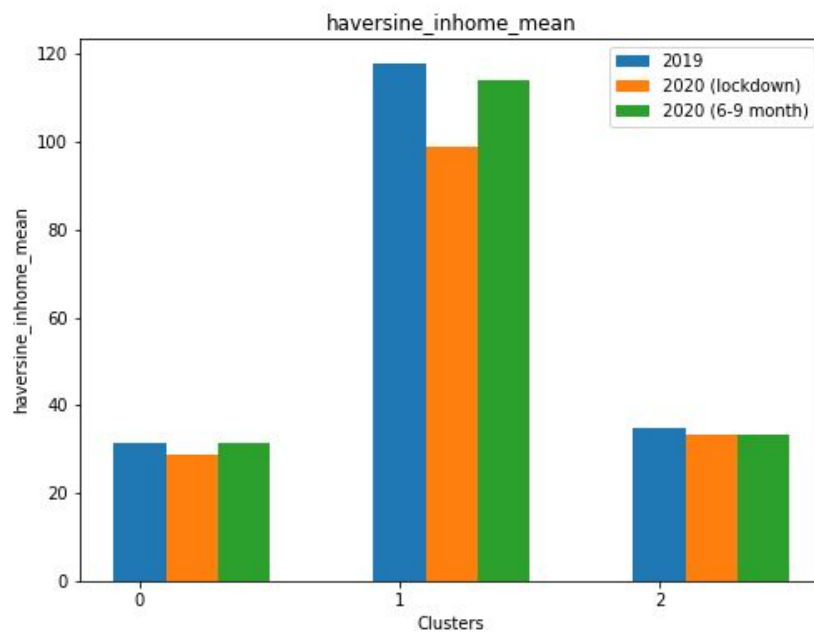
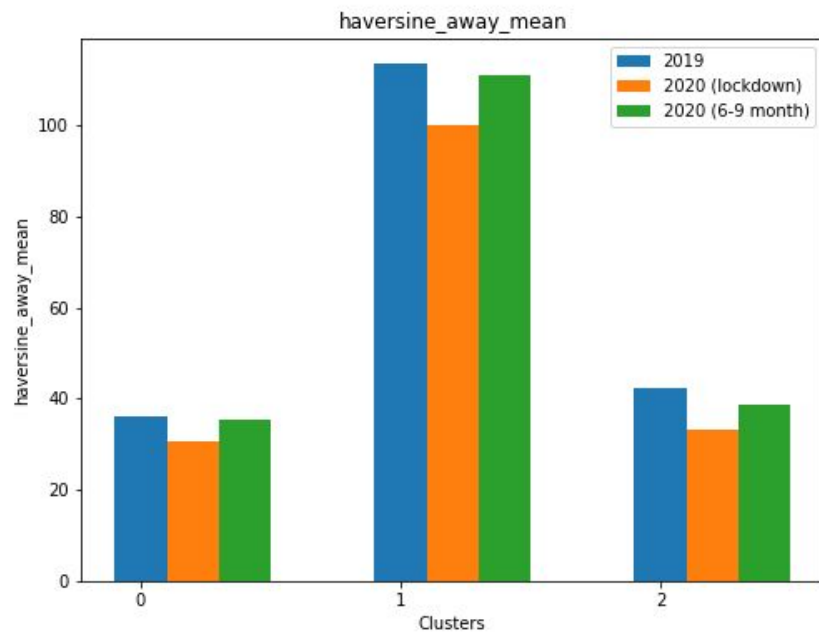
Порівняння та динаміка ●●

Кластер 1 під час та після локдауну майже не змінив кількість переміщень

Інші кластери повноцінно не відновили кількість переміщень 2019 року



Порівняння та динаміка ●●



Задача 1: підсумки ●●

Кластери:

1. Робочий клас із достатком вище середнього
2. Любителі міжобласних переміщень
3. Середній клас без особливого відпочинку

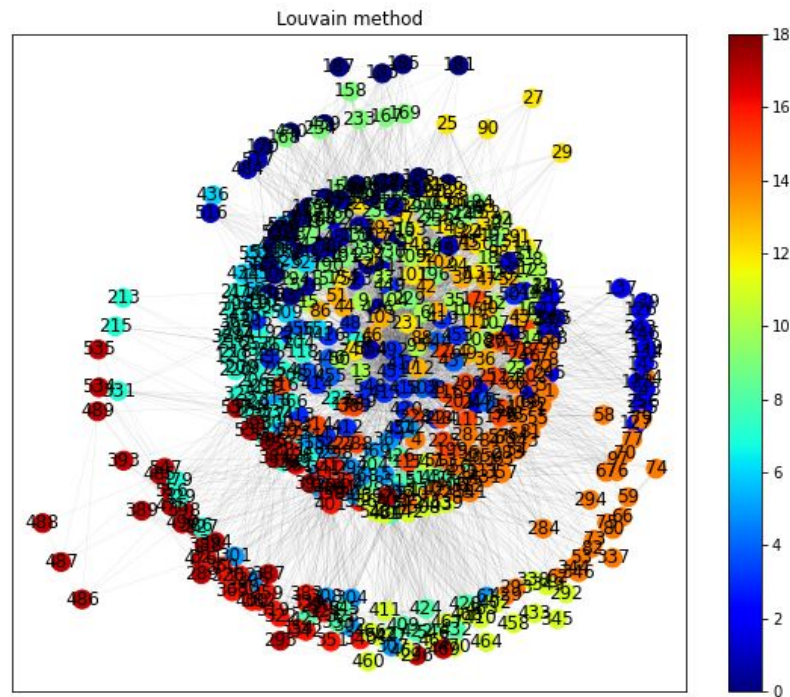
Карантин:

1. Після локдауну **мобільність повернулася** (відновлення роботи + транспорт)
2. Всього **лише 2%** від тих, хто почав зменшувати мобільність, продовжив після локдауну
3. Робочі переміщення повернулися в колишнє русло **(-3%)**

Задача 1: підсумки ●●

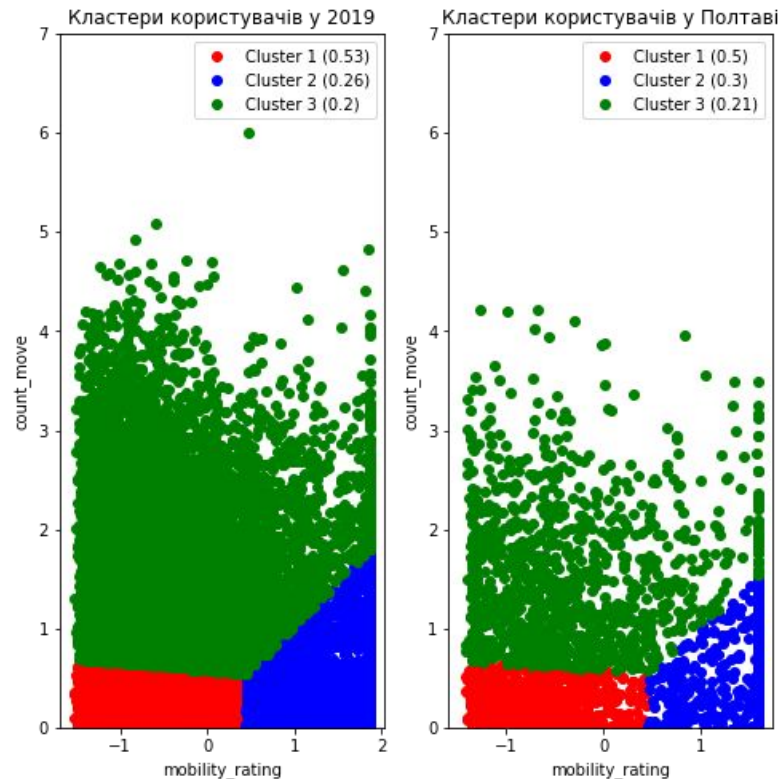
Як покращити:

- Довести до фіналу реалізацію **за допомогою графів**
- **Проріджуємо** ініціалізований граф: видаляємо ребра серед малозв'язних вершин
- **Метод Лувена** для знаходження спільнот вершин (найбільш тісно пов'язані між собою)



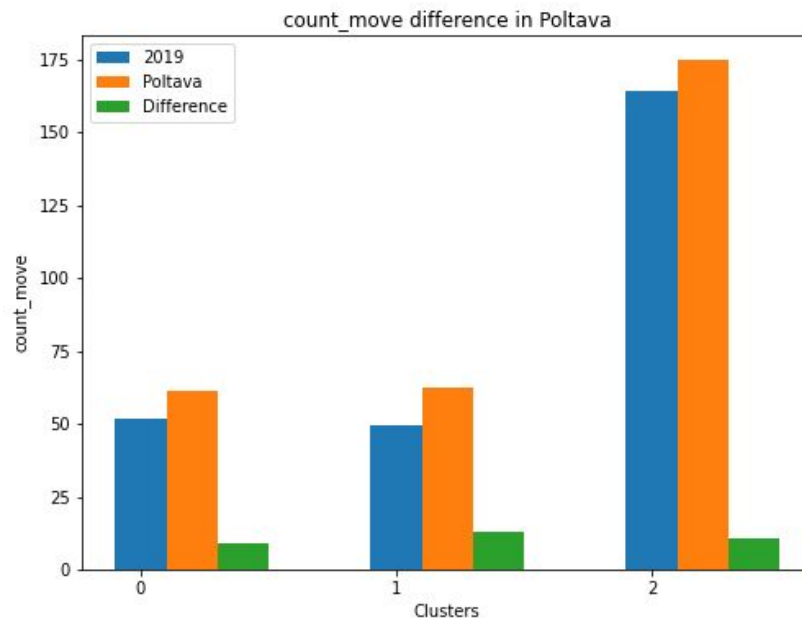
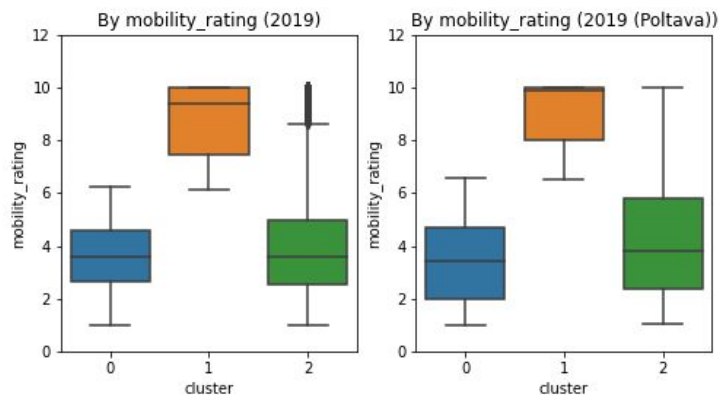
Завдання 2.1: порівняння сегментів ●●

- **Дано:** Датасет переміщень за 2019 та транзакційні дані Полтави
- Візуально **межі кластерів залишились незмінними**
- Але два кластери мають **меншу густину**
- Всі точки Полтави **на 10% ближче** до своїх центроїдів



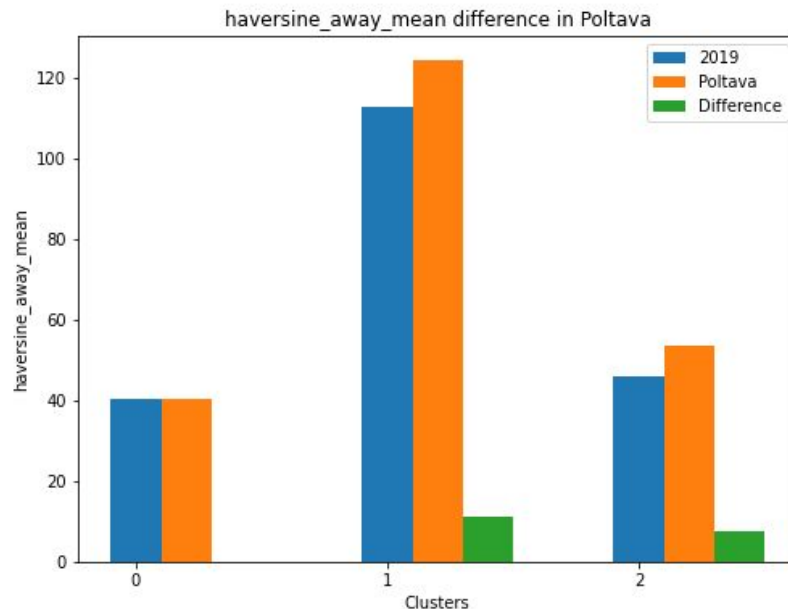
Порівняння по індикаторам ●●

- Рівень мобільності майже без змін



Порівняння по індикаторам ●●

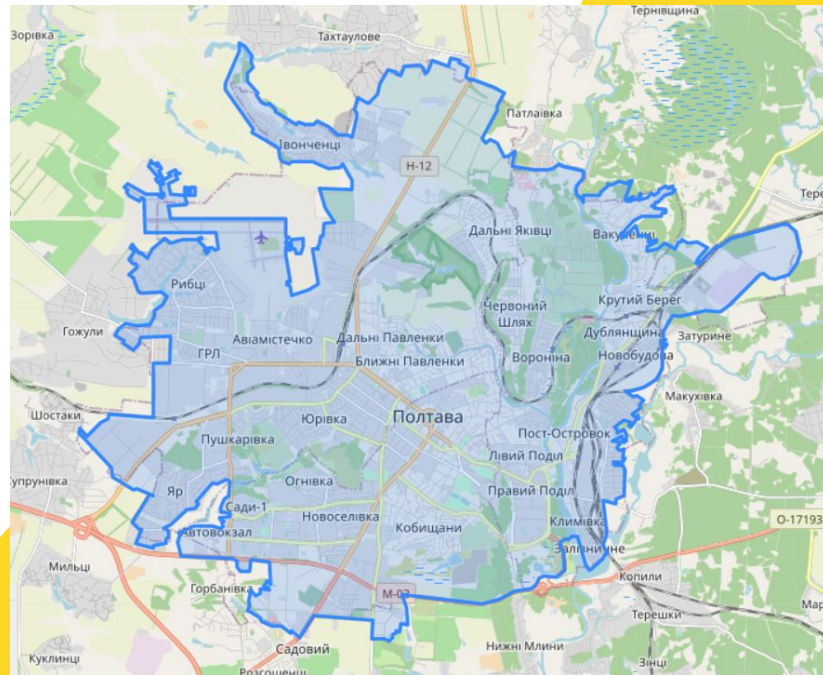
- У так званих “любителів дальніх подорожей” спостерігається **збільшення середньої відстані** переміщення у непомашні райони
- Можливо, це пов'язане з тим, що серед популярних напрямів є Київ, Харків, Дніпро



Завдання 2.2: пошук точок в'їзду/виїзду

Ключові фактори при пошуку:

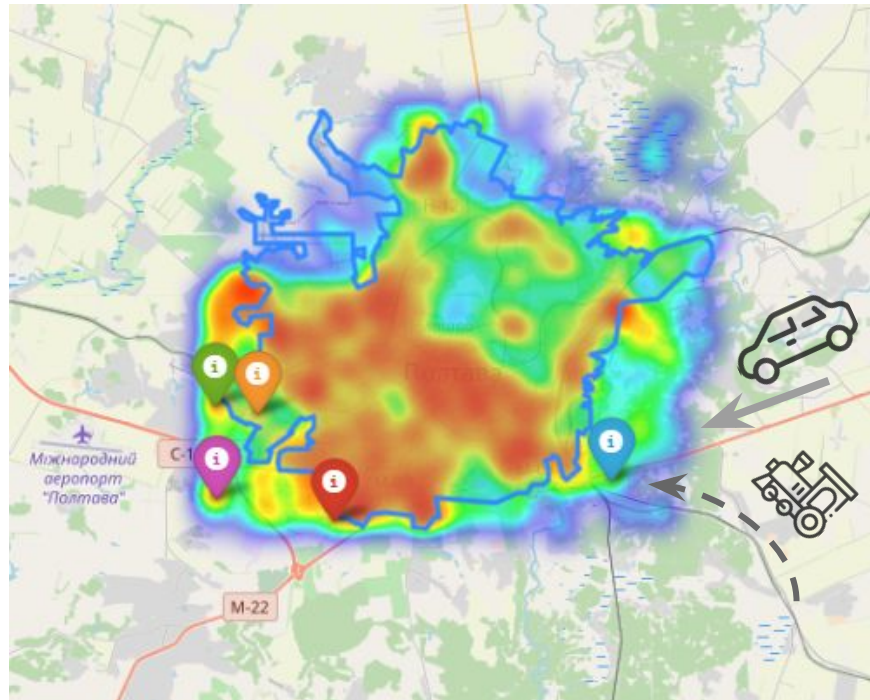
- Кількість чек-інів у точці
- Кількість унікальних користувачів
- Наскільки часто користувачі кластеру взагалі бувають в Полтаві



Кластер 0 ●●

село Копили (9к перетинали,
30% часу у Полтаві)

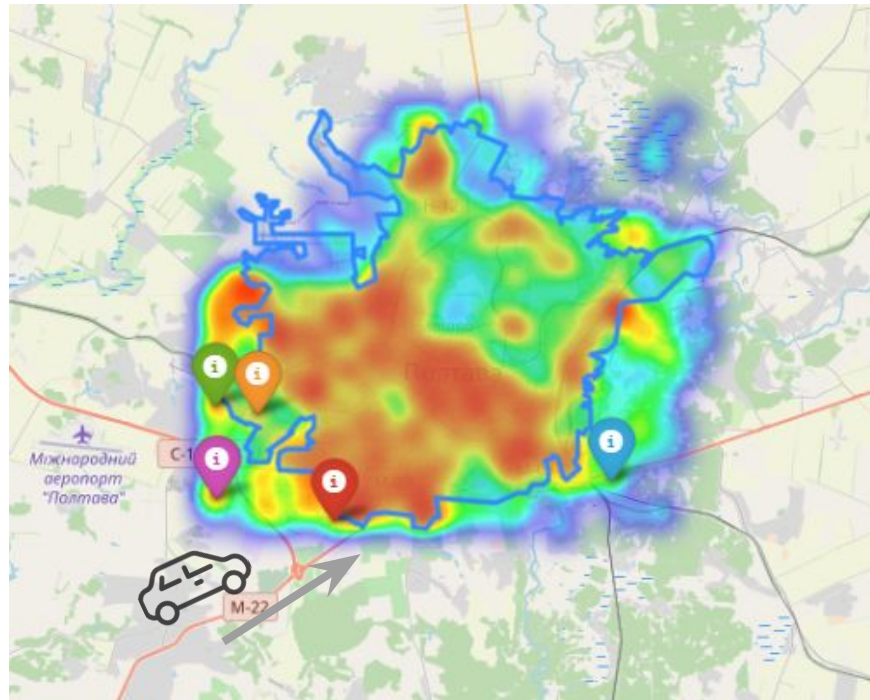
1. **Посадковий пункт "3 км"**
(характерні скупчення по
маршруту електрички
Полтава-Лозова 47 зупинок)
2. **Розв'язка траси М-03 (під'їзд
до Полтави).** 40% транзитом
Харків - Київ.



Кластер 0 ●●

село Розсошенці (6к, 34% у
Полтаві)

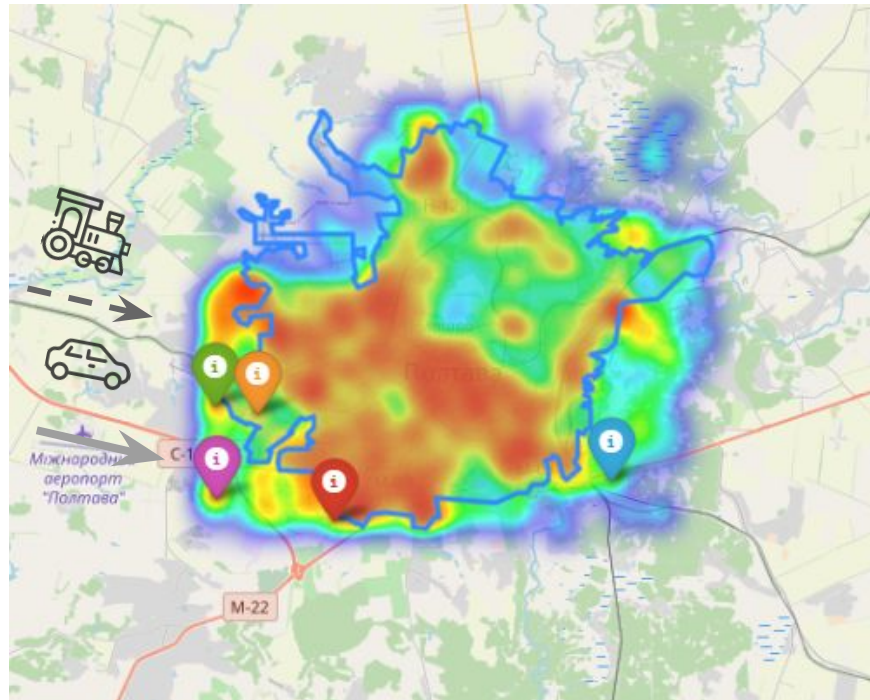
1. **Розв'язка М-22 та М-03**
(велика активність після
цього в'їзду на автовокзалі).
Маршрут до Кременчука.



Кластер 0 ●●

село Супрунівка (8.5к
перетинали, 22% часу у Полтаві)

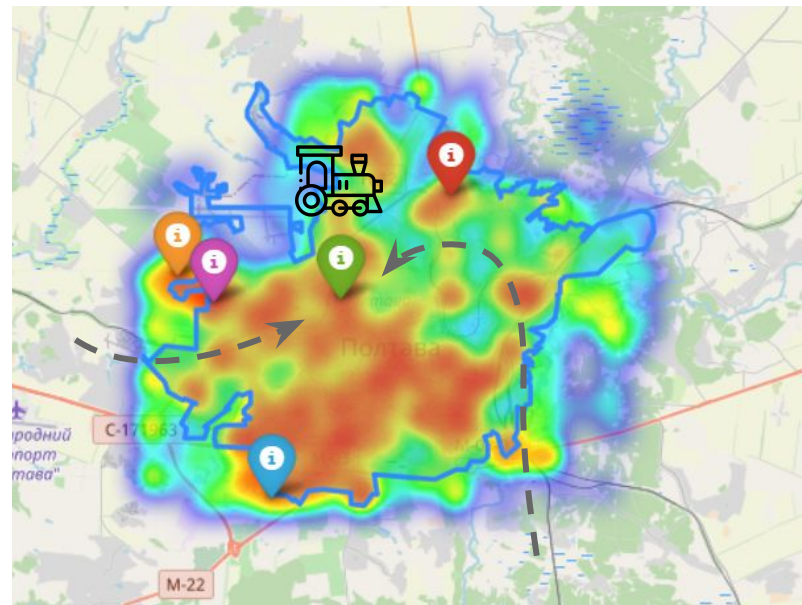
1. **Посадковий пункт**
"Супрунівка" (великі станції:
Миргород, Лубни, Ніжин)
2. **М-03**. 25% їдуть із Західної
України.



Кластер 1 ●●

Посадковий пункт Полтава - Київська

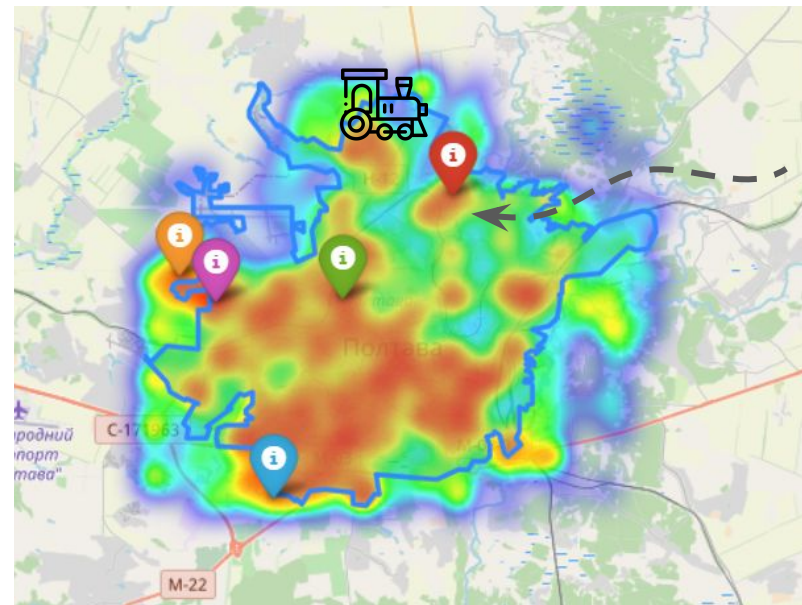
- Центральна станція **(найбільше їдуть із Миргорода) Серед великих напрямів:** Київ, Харків, Івано-Франківськ, Кіровоград.
- Підтвердження нашим міркуванням знайдені серед напрямів УкрЗалізниці



Кластер 1 ●●

Посадковий пункт Дублянщина

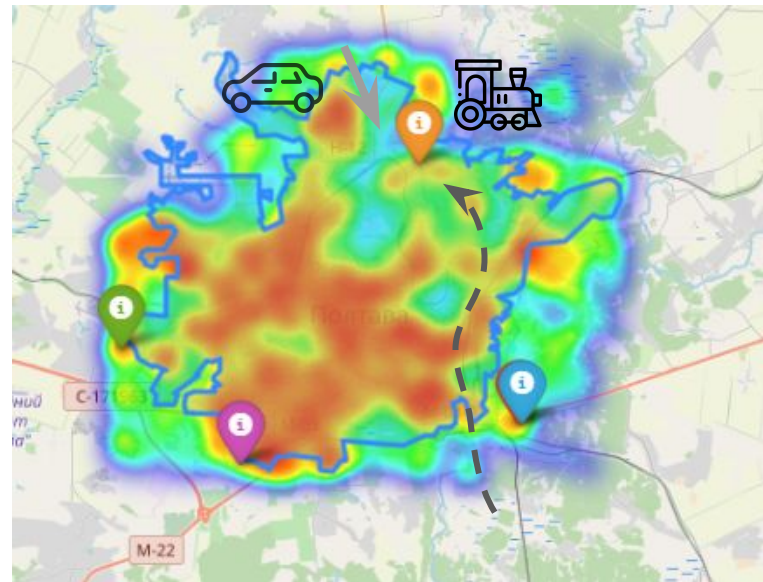
- Маршрут через села, паралельні до М-03, майже до Харкова (село Огульці)



Кластер 2 ●●

село Ялівці

- посадковий пункт Ялівці
- траса Н-12, на Суми (дуже рідко приїжджають саме до Полтави, переважно транзит)



Завдання 2.3: мета поїздки

Труднощі:

- “Дивні” користувачі. **Рішення:** зробити прохвилинний кроп
- Дефіцит змістовних ознак
- З якого боку підійти?

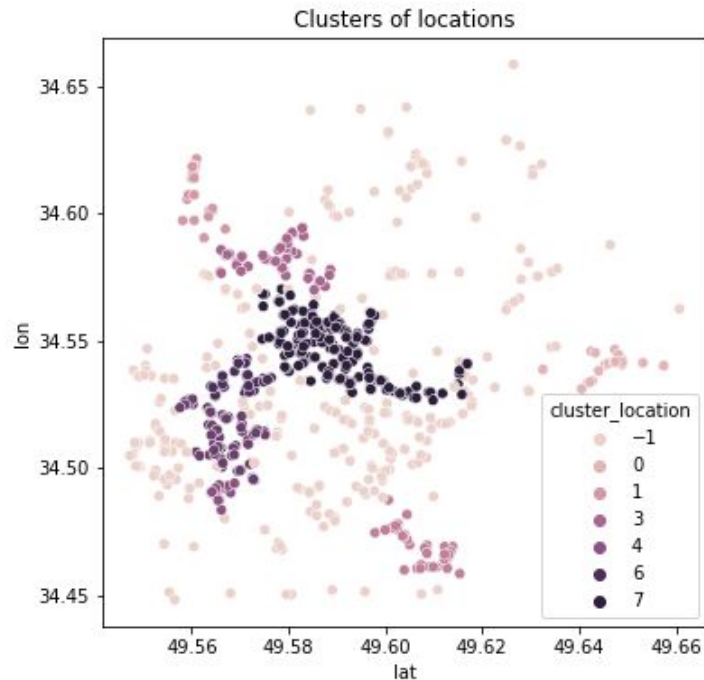
Рішення: кластеризація по щільності локацій

	user_id	event_dt	lat	lon
6555845	134897	2020-09-01 00:09:37	49.538216	34.516945
6548321	134897	2020-09-01 00:39:37	49.538216	34.516945
6556240	134897	2020-09-01 01:09:37	49.538216	34.516945
6554889	134897	2020-09-01 01:39:38	49.538216	34.516945
6548561	134897	2020-09-01 02:09:38	49.538216	34.516945

6554521	134897	2020-09-13 13:05:38	49.603138	34.529480
6549983	134897	2020-09-13 13:05:39	49.603138	34.529480
6554046	134897	2020-09-13 13:16:58	49.600334	34.530636
6558177	134897	2020-09-13 13:16:59	49.600334	34.530636
6557437	134897	2020-09-13 13:17:15	49.604733	34.530014
6552714	134897	2020-09-13 13:17:16	49.604733	34.530014

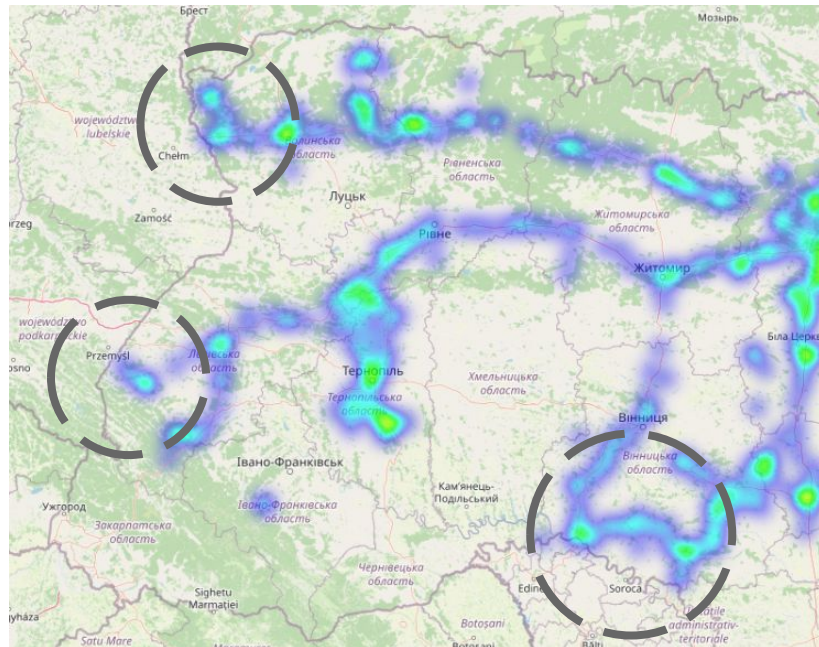
Кластеризація локацій ●●

- Прогнозуються **кластери із різними щільностями** та нестандартної форми - **обрали HDBSCAN**
- Налаштування на невелику кількість кластерів, аби отримати **узагальнюючу інформацію про локації та оточення**



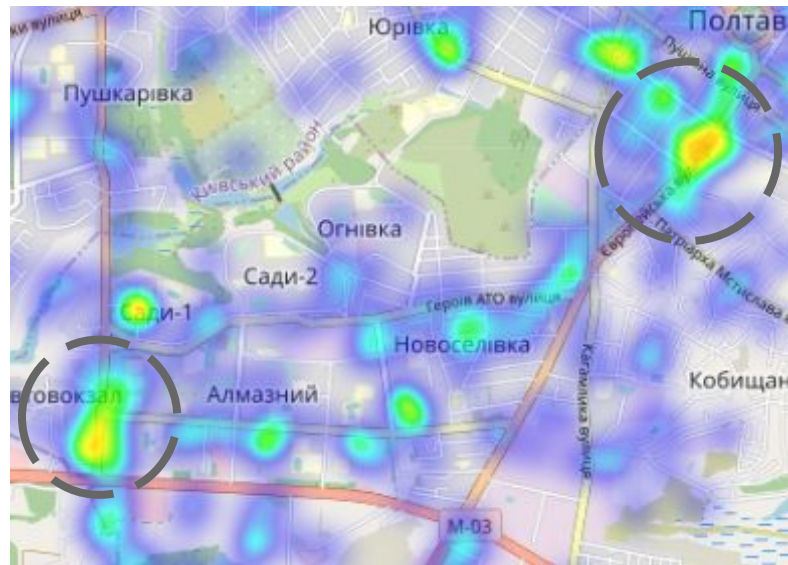
Отримані інсайти ●●

Для кластеру 1 (любители далеких подорожей) на вихідних помічена **активність у прикордонних містах**, яка веде до Полтави через Київ (можливо, **підприємницька діяльність**)



Отримані інсайти ●●

Частина представників кластеру 2 (майже безвиїзні групи) на вихідних приїжджає у Полтаву на автобусі **(активність на автовокзалі)** і через декілька годин помічені на Центральному ринку **(з метою закупитися продуктами або** навпаки, **продавати щось, оскільки їдуть із сіл)**

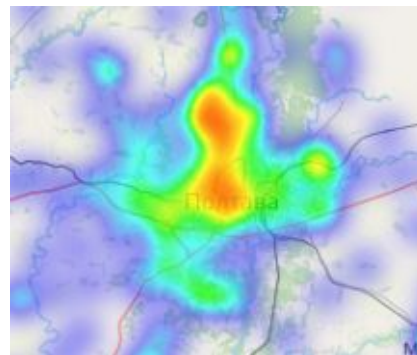


Отримані інсайти ●●

70% користувачів кластеру 0 демонструють помітну різницю по транзакціям на період: **будні дні, з 7:00 по 21:00** та відповідно весь інший час.

Із 21:00 по 7:00 збільшується активність в області: Красноград, Миргород, Нові Санжари, Диканька.

Отже, із великою ймовірністю - **це трудова міграція.**



Отримані інсайти ●●

Для більшості користувачів, незалежно від кластеру, характерним паттерном на вихідних є прогулянка по центру Полтави (за умови, що це не транзитне переміщення).

Мета: **відпочинок та культурний розвиток.**



Задача 2: підсумки ●●

Як покращити:

- Використати інформацію з відкритих джерел: **OpenStreetMap** (тип локації, наближчі дороги)
- Більш детально опрацювати інформацію **про дивну активність користувачів**
- Спробувати **математичні методи** вирішення задачі: раніше запропоновані графи або ланцюги Маркова

Дякуємо за увагу!