Пал Бодай

Советник Президента Центрального статистического управления Венгрии

Виртуальная Венгрия

Взаимосвязь статистических и административных баз данных

Проблема не в нехватке данных — проблема в отсутствии интеграции

Разница традиционных и административных данных

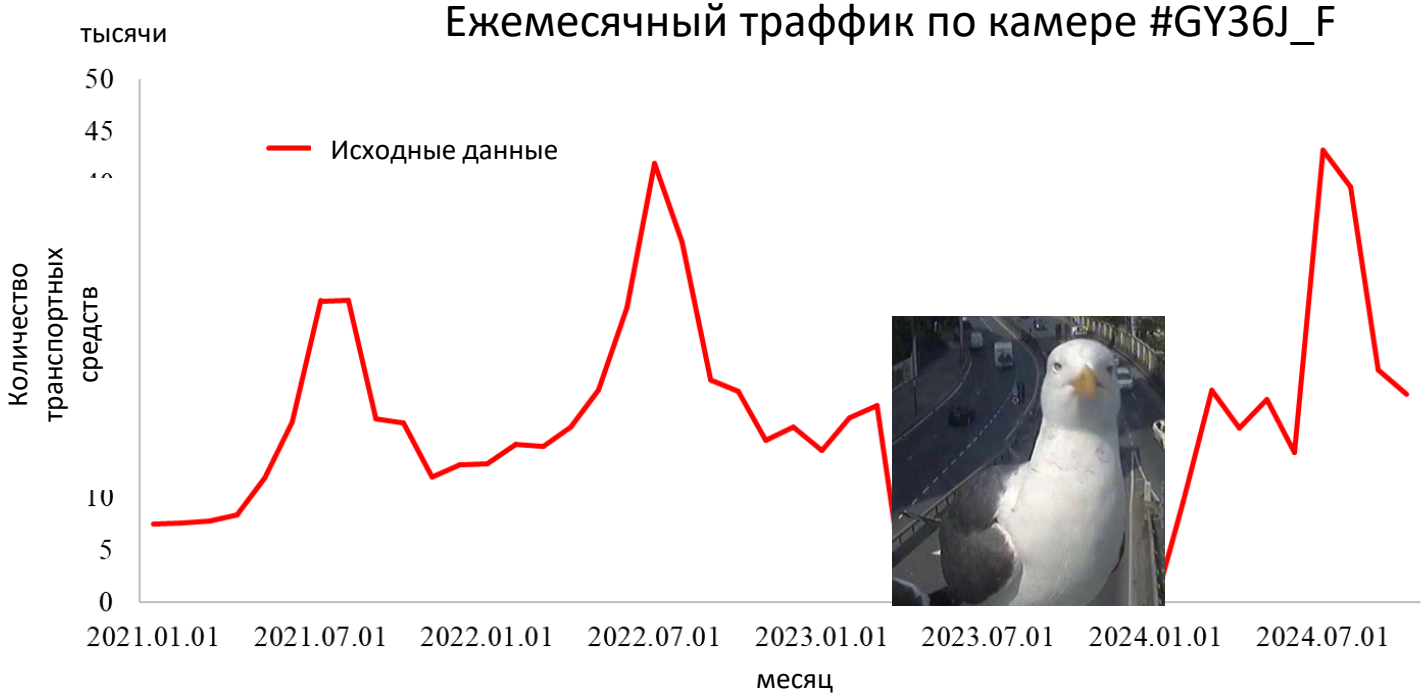
Традиционные данные	Административные данные
Плюсы:	**Плюсы:**
• Разрабатываются для статистических целей	• Масштабность
• Высокий уровень методологического контроля	• Детализация
• Международная сопоставимость	• Регулярное обновление
Минусы:	**Минусы:**
• Дорого	• Не предназначены для статистики
• Медленно	• Проблемы с качеством
• Ограниченная степень детализации	• Разрозненность

ИИ в официальной статистике: от инструмента к возможностям

- Данные в избытке – проблема в их интеграции
- Административные и новые источники данных - неоднородны
- **Что позволяет ИИ:**
  - Очистку данных и повышение качества
  - Сопоставление и объединение данных из разных систем
  - Заполнение пропусков и импутация
  - Автоматизацию процессов
  - Выявление закономерностей в сложных данных
- Новый подход: от производства данных – к их интеграции

Оценка трафика на пограничных пунктах с использованием машинного обучения

- Ежемесячная агрегация данных с камер на пограничных переходах
- Заполнение пропущенных значений методом интерполяции

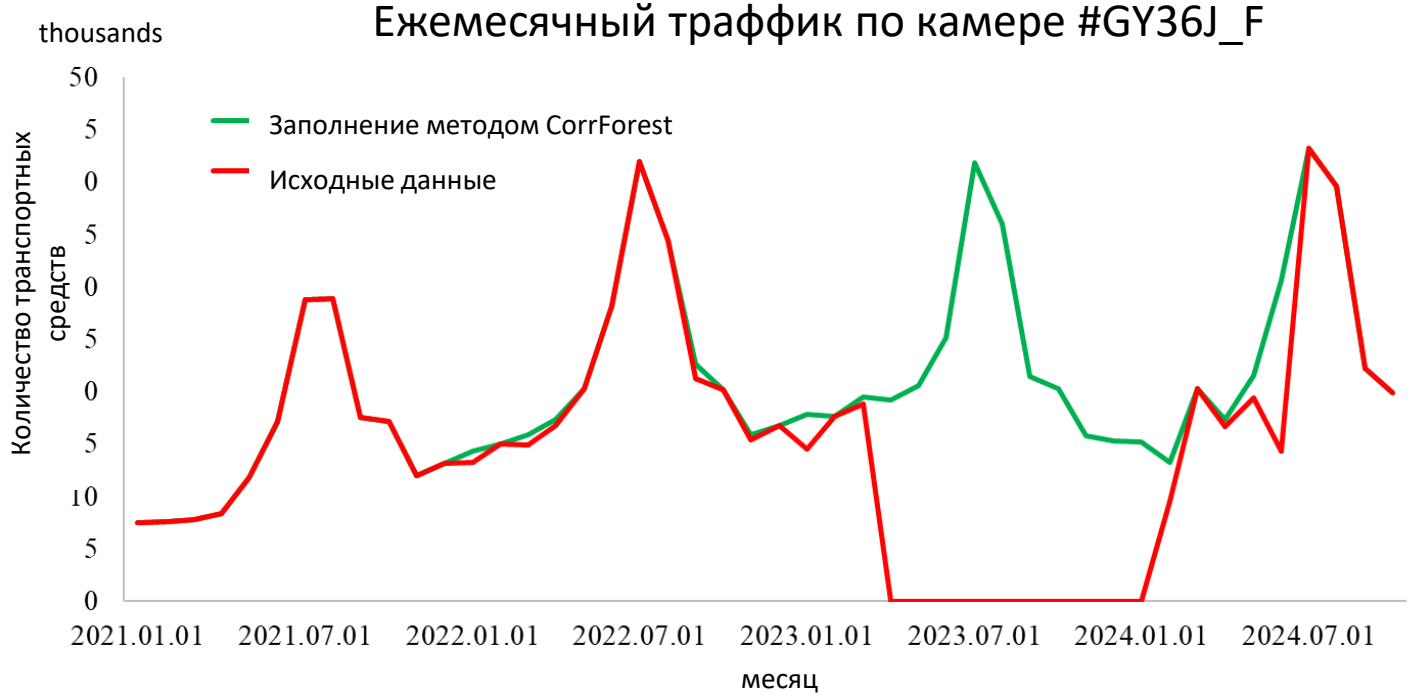


$$\mathrm{I}_G(p) = \sum_{i=1}^{J}\left(p_i \sum_{k\neq i} p_k\right) = \sum_{i=1}^{J} p_i(1-p_i) = \sum_{i=1}^{J}(p_i - p_i^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2 = 1 - \sum_{i=1}^{J} p_i^2.$$

$$\Delta\sigma^2 = \sigma_p^2 - \left(\frac{n_l}{n_t}\sigma_l^2 + \frac{n_r}{n_t}\sigma_r^2\right)$$

Оценка трафика на пограничных пунктах с использованием машинного обучения

- Ежемесячная агрегация данных с камер на пограничных переходах
- Заполнение пропущенных значений методом интерполяции



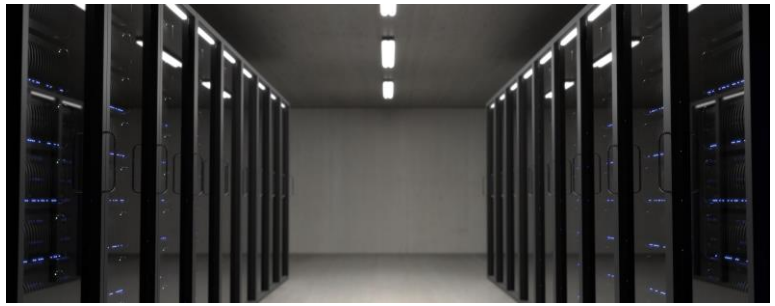
$$\mathrm{I}_G(p) = \sum_{i=1}^{J}\left(p_i \sum_{k\neq i} p_k\right) = \sum_{i=1}^{J} p_i(1-p_i) = \sum_{i=1}^{J}(p_i - p_i^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2 = 1 - \sum_{i=1}^{J} p_i^2.$$

$$\Delta\sigma^2 = \sigma_p^2 - \left(\frac{n_l}{n_t}\sigma_l^2 + \frac{n_r}{n_t}\sigma_r^2\right)$$

Проект Венгрия онлайн (VIMA)

Отдельные базы данных

- censuses (population census)
- Административные данные (налоговые органы, социальные службы)
- регистры (бизнес регистры)
-крупномасштабные выборочные обследования (обследование рабочей силы, доходов и условий жизни)



Анонимизация

Объединение

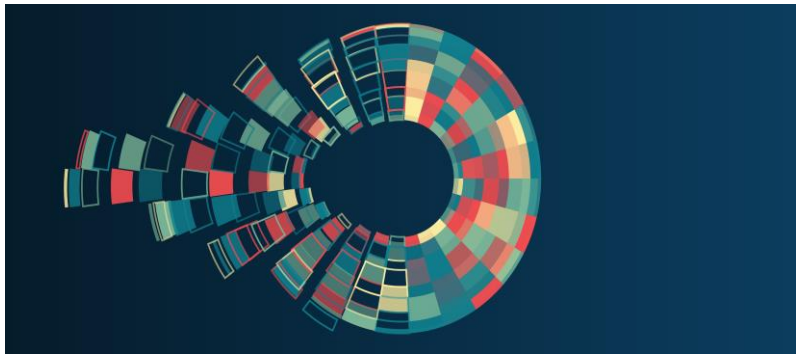
Импутация, экстраполяция

Современные методы:

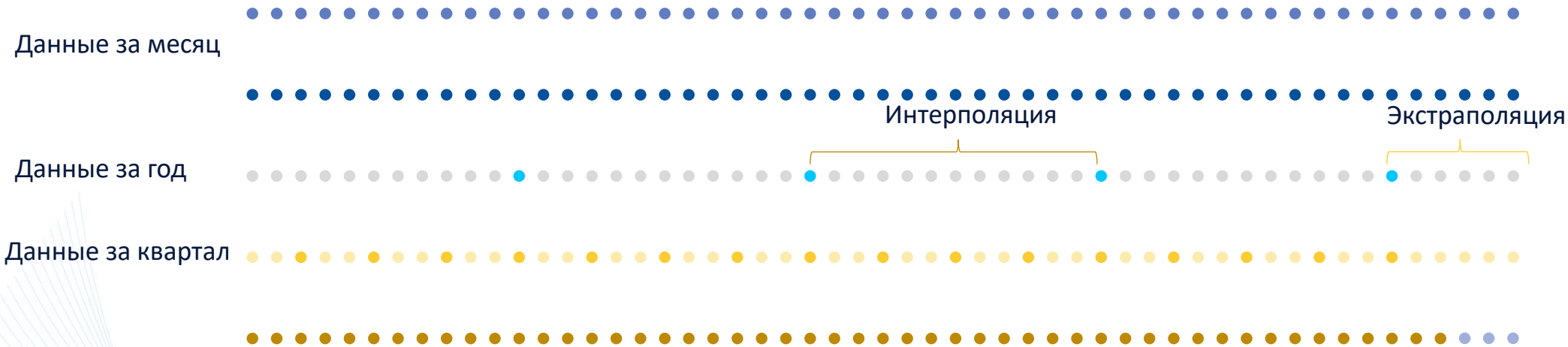
Статистическое сопоставление, Автоматизация, микросимуляция, машинное обучение, глубокое обучение

Объединенные базы данных

- обезличенные базы даных
- индивидуальная или низкоуровневая агрегация
- постоянно обновляемые данные

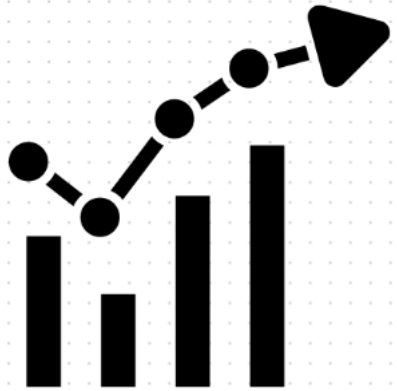


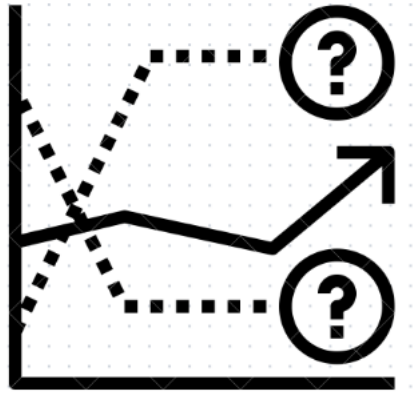
Импутация


Данные за месяц
Интерполяция
Экстраполяция
Данные за год
Данные за квартал

Ключевые задачи

- Актуальность
- Детализация
- Прогнозируемость
- Аналих сценария





Благодаря объединению баз данных можно получить значительно больше информации, чем сумма сведений, содержащихся в исходных базах данных.

Перепись населения

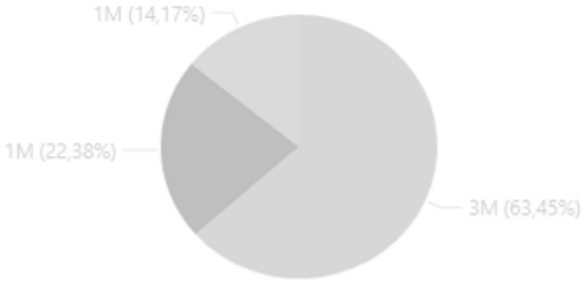
- **Использованные источники данных:**
  - **Перепись населения 2022**
  - **Налоговые данные по заработной плате – налоговые органы**

Что мы знали «вчера»?

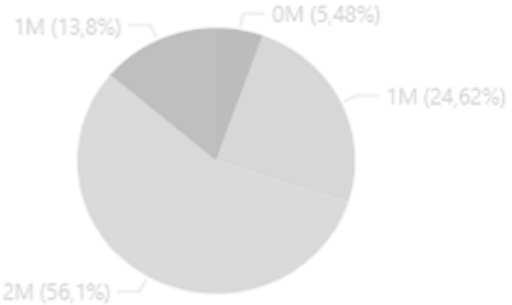
Перепись населения

Уровень образования

commuting

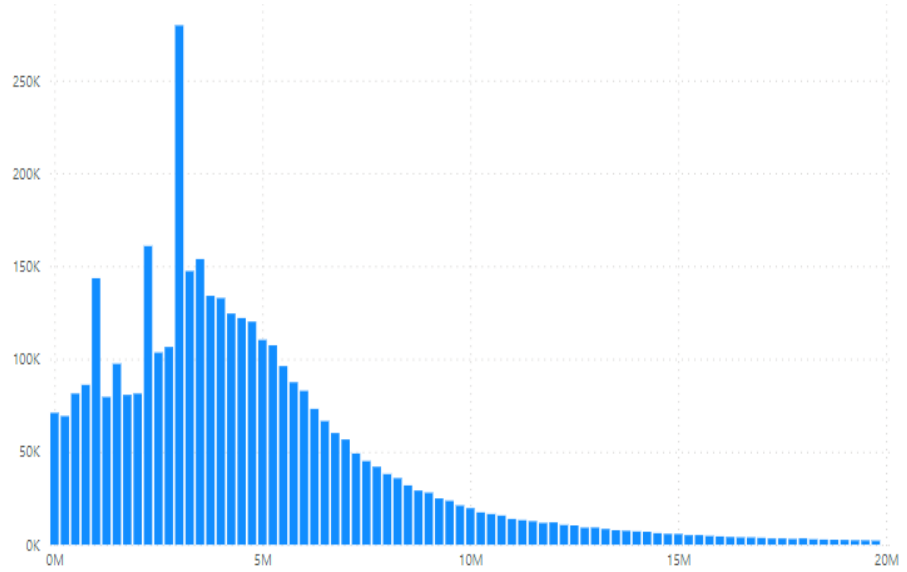


digital skills



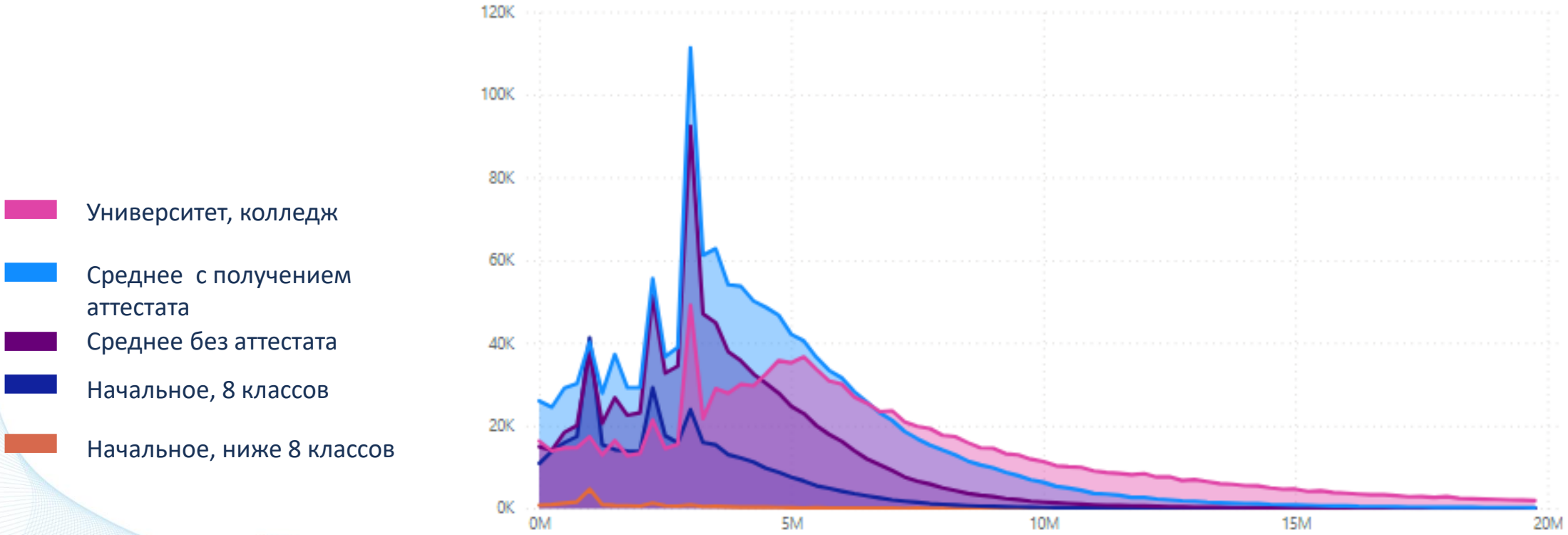
Налоговые данные

Средняя заработная плата

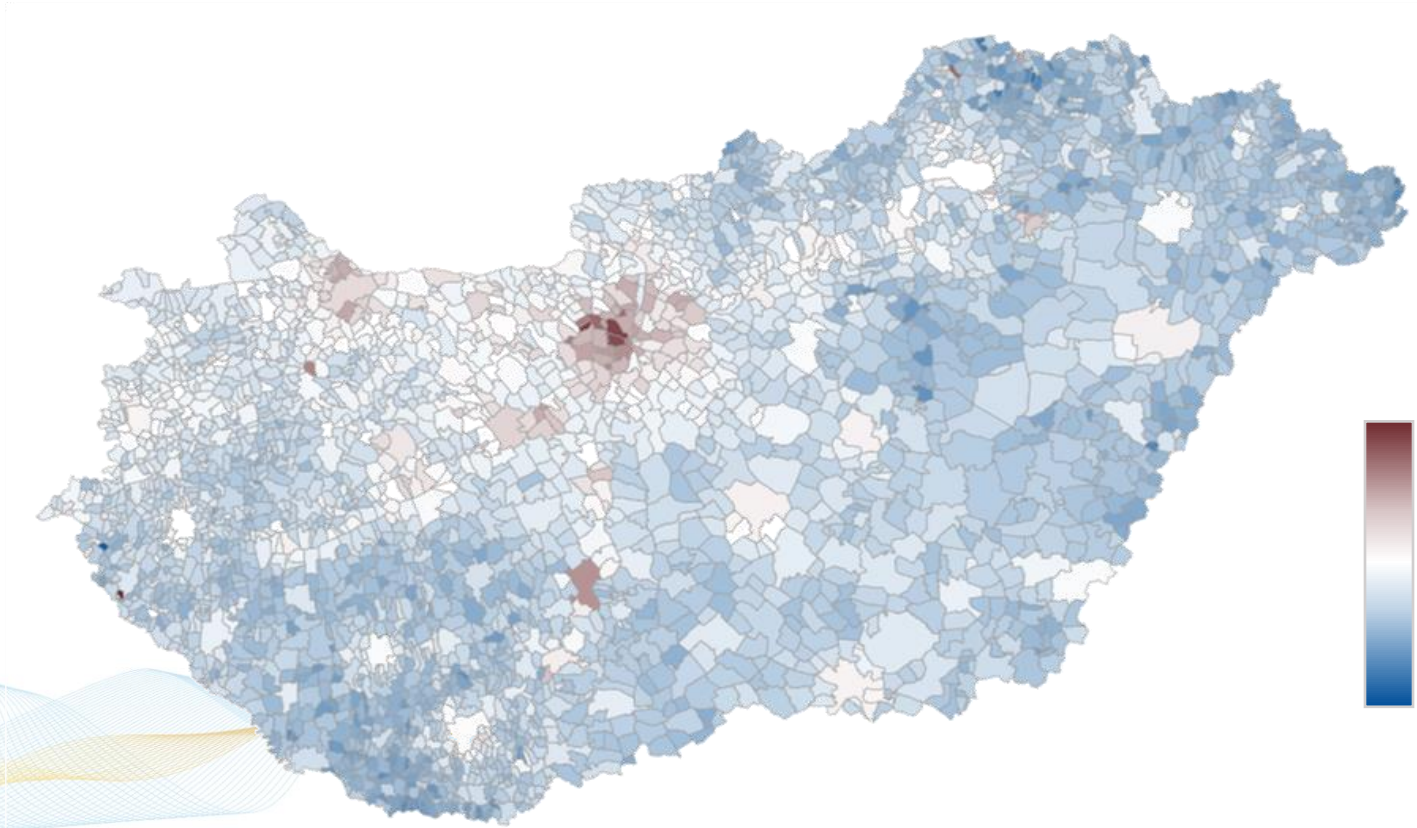


Что мы знаем «сегодня»

Распределение заработной платы по уровню образования



Годовая заработная плата по населенным пунктам



10млн форринтов (HUF)

5M млн форринтов (HUF)

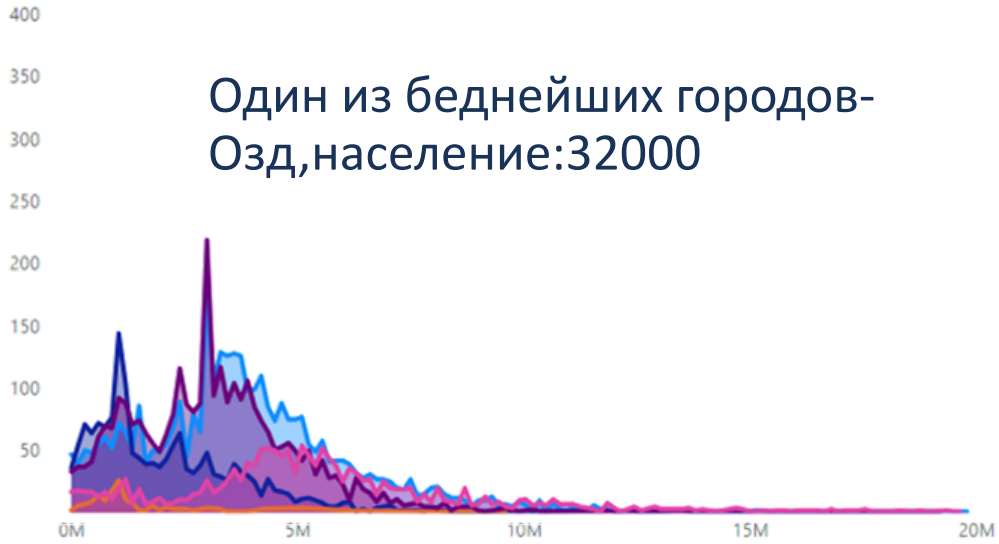
0M млн форринтов (HUF)

HUNGARIAN CENTRAL STATISTICAL OFFICE

…по уровню образования

Один из состоятельных городов – Сентендре, население: 26 000)

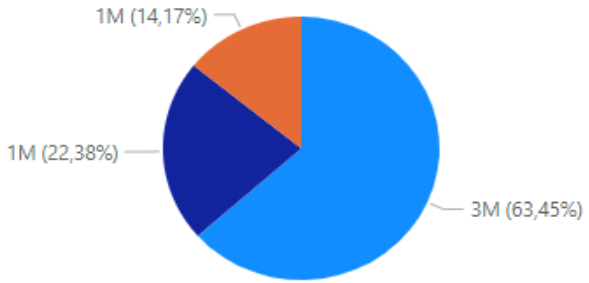
Один из беднейших городов- Озд,население:32000



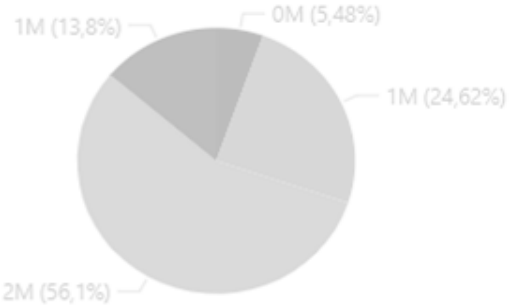
Что мы знали «вчера»

Перепись населения

Маятниковая миграция

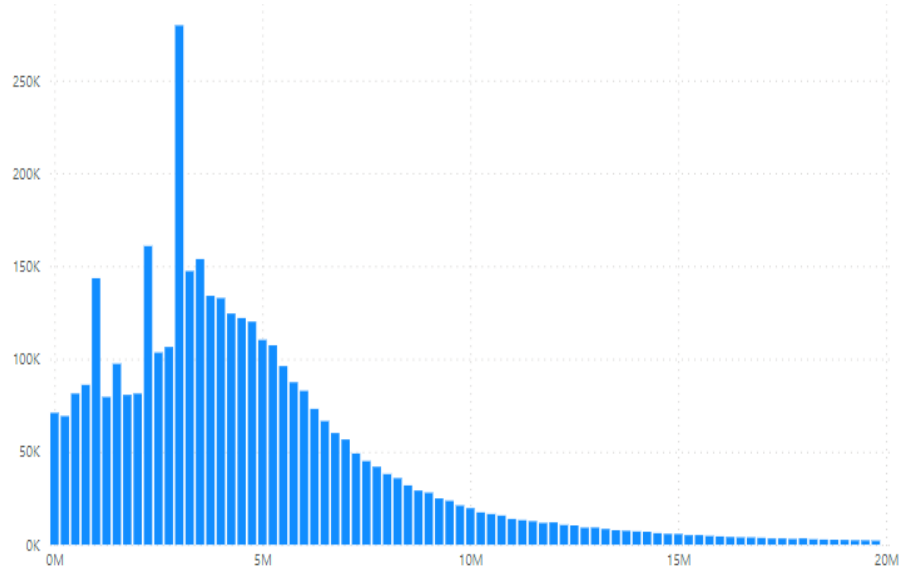


Цифровые навыки

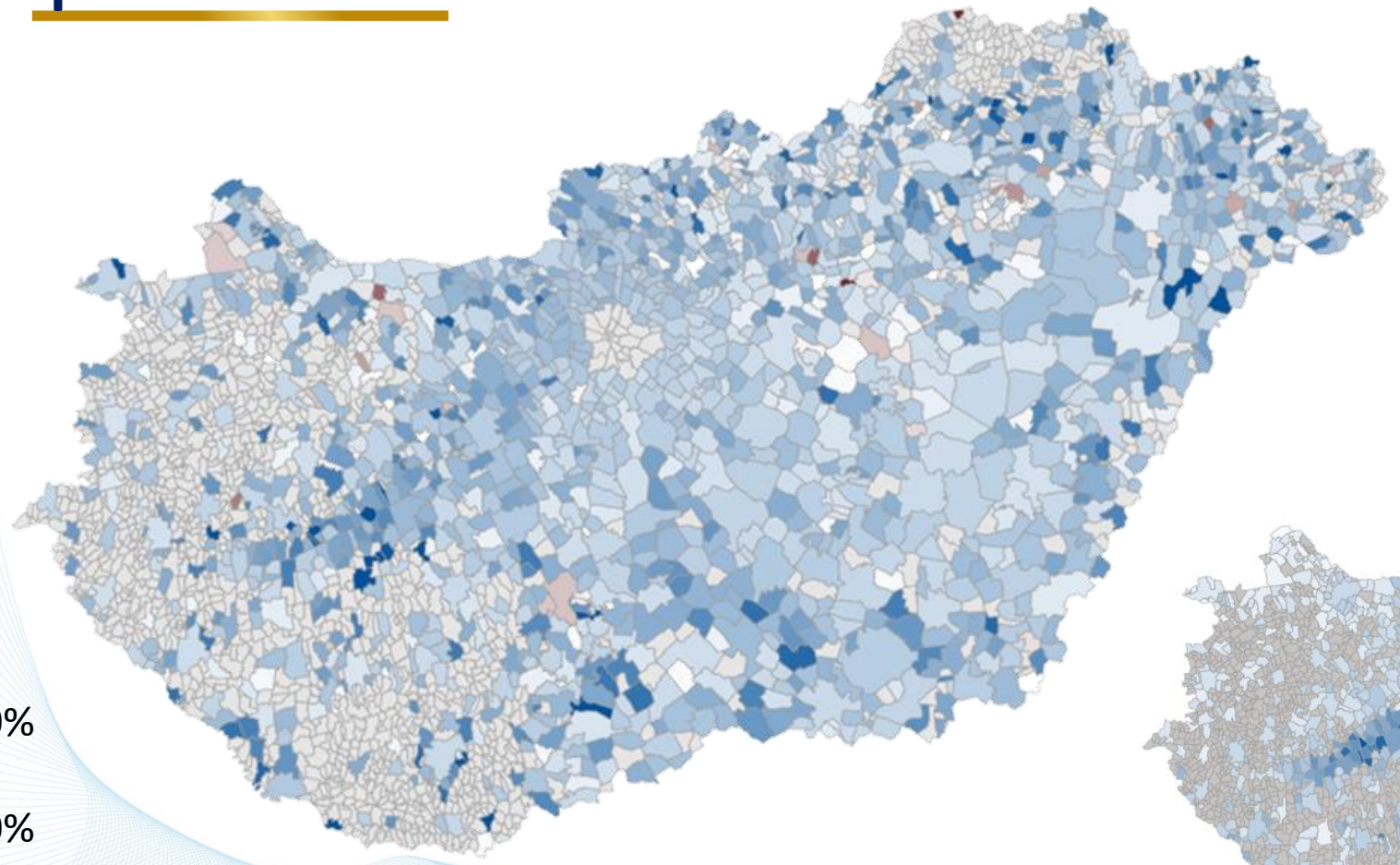


Налоговые данные

Средняя заработная плата

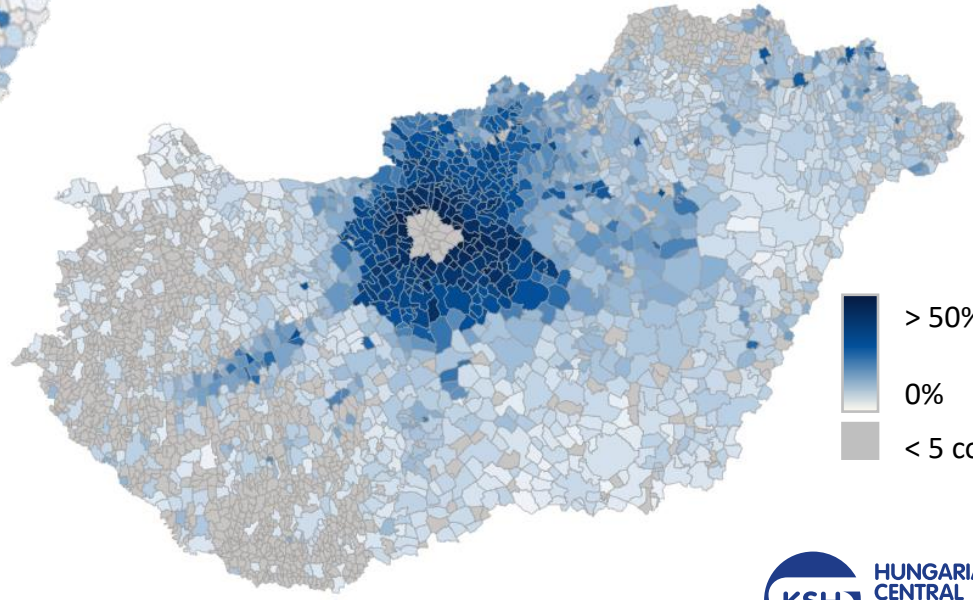


Заработная плата работников, совершающих поездки в Будапешт, по сравнению с заработной платой местных работников



300%
100%
20%

Доля лиц, совершающих поездки в Будапешт



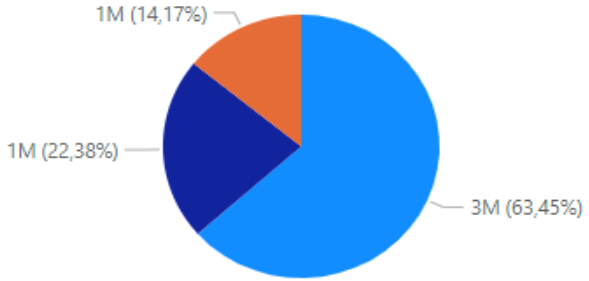
> 50%
0%
< 5 commuters

Что мы знали«вчера»

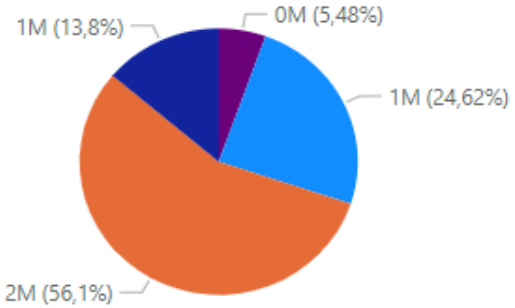
Перепись населения

Уровень образования

Маятниковая миграция

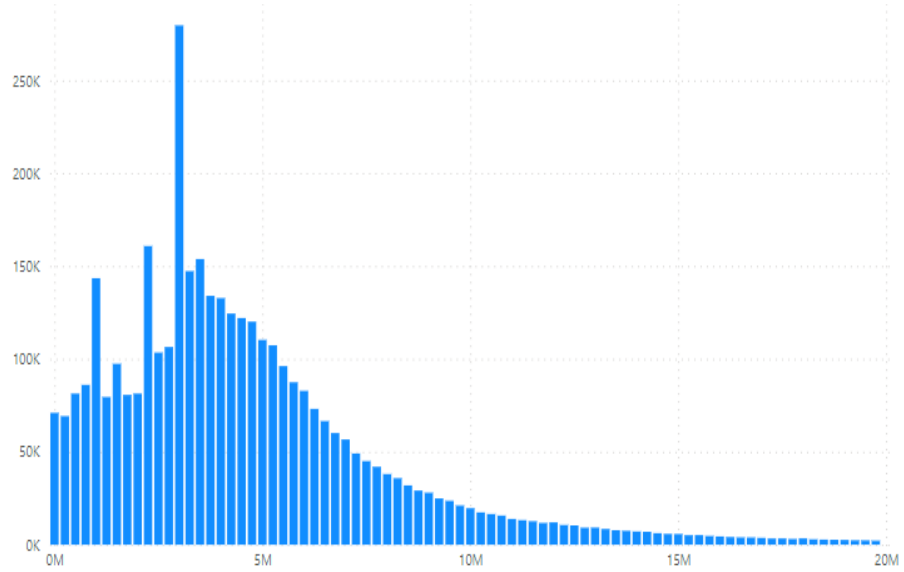


Цифровые навыки

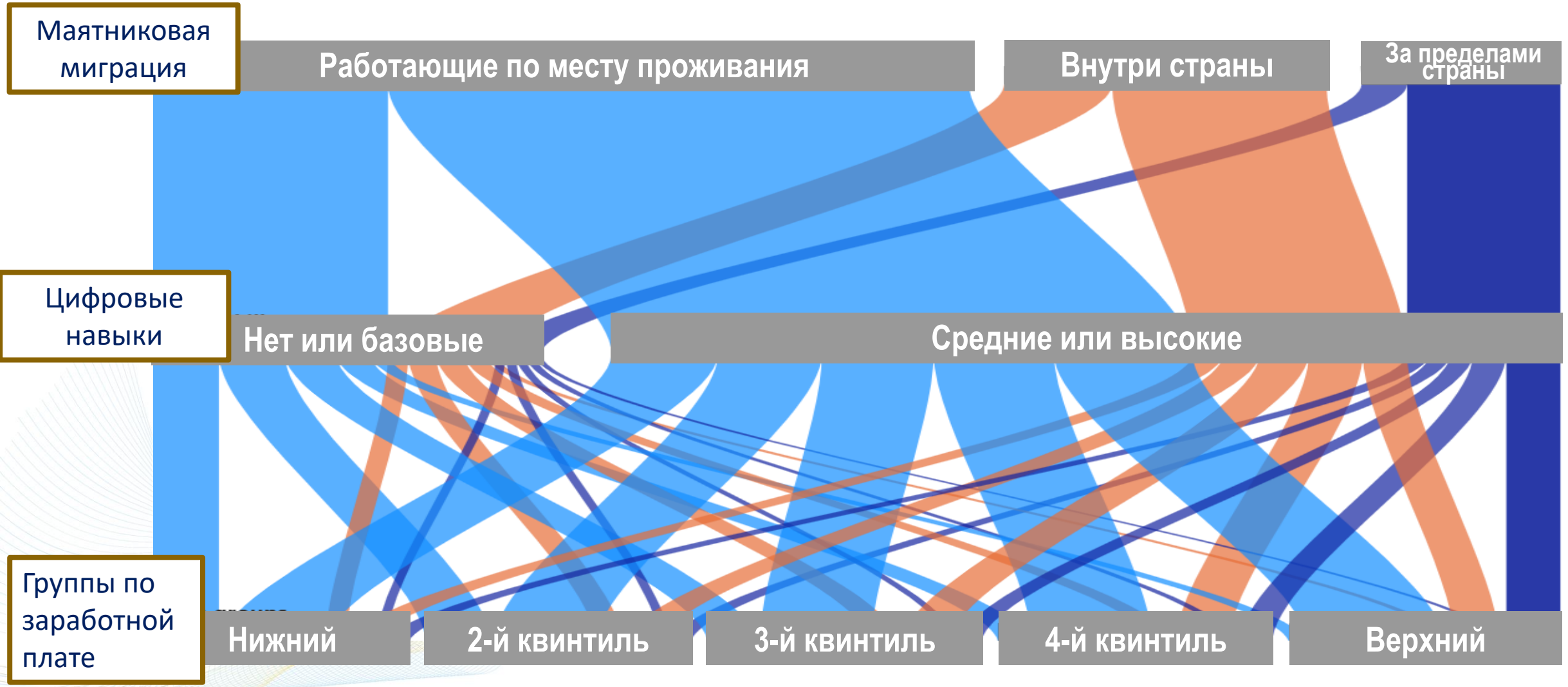


Налоговые данные

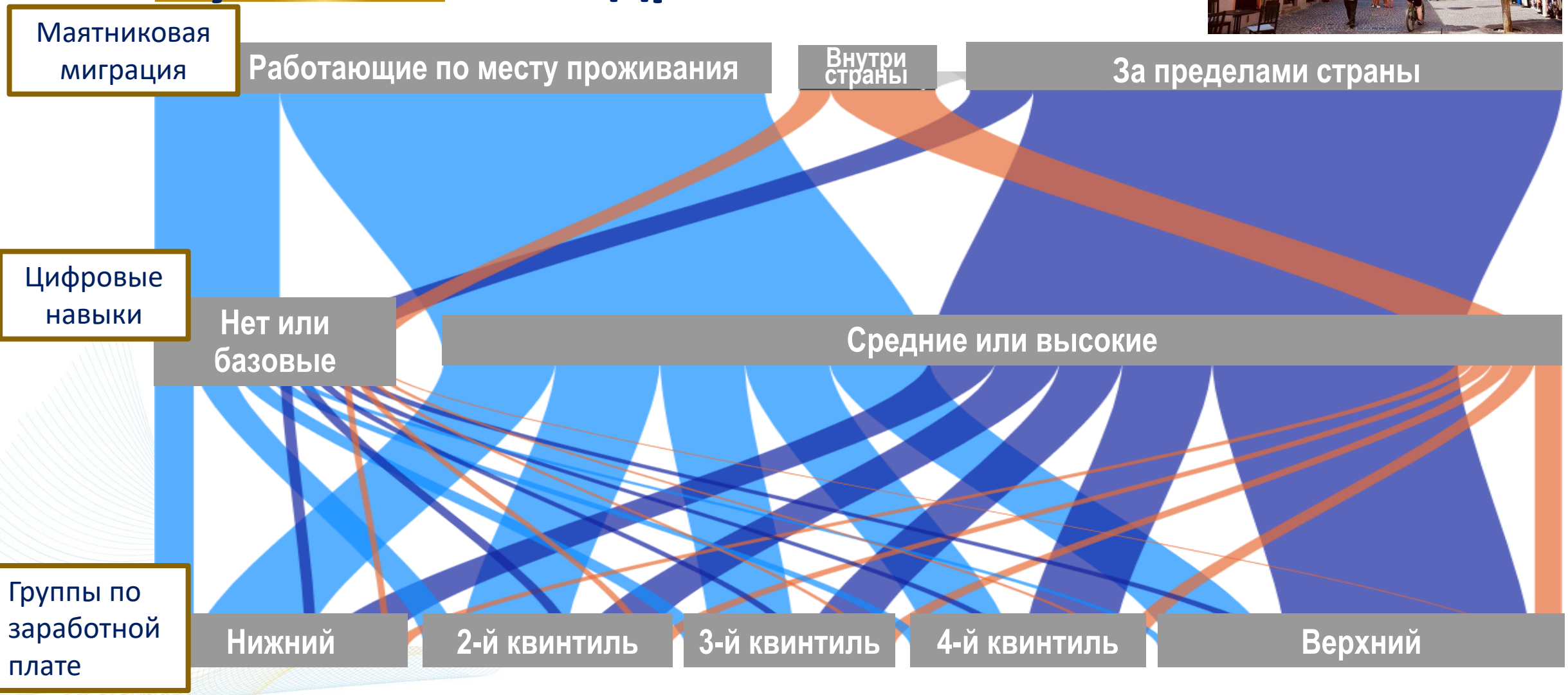
Средняя заработная плата



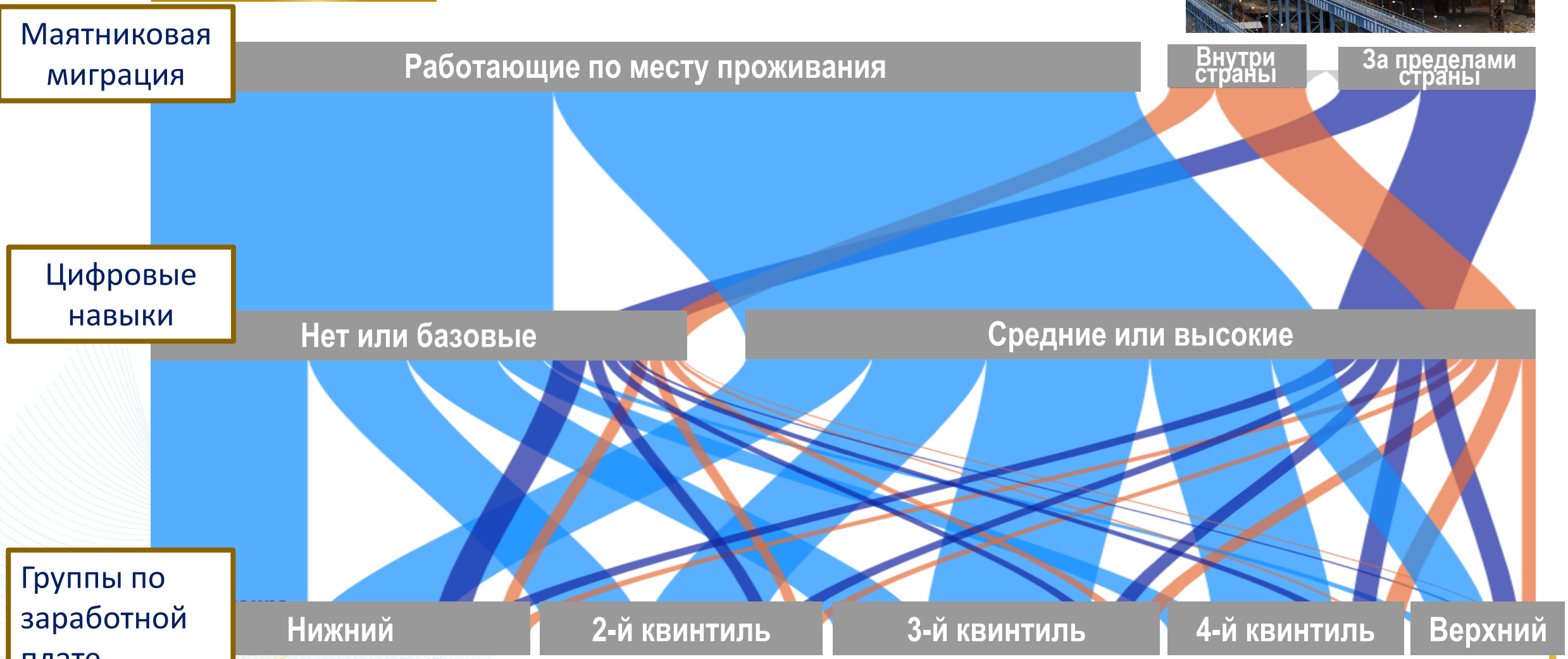
Взаимосвязь переменных из разных баз данных



…для каждого отдельного населенного пункта: Сентендре

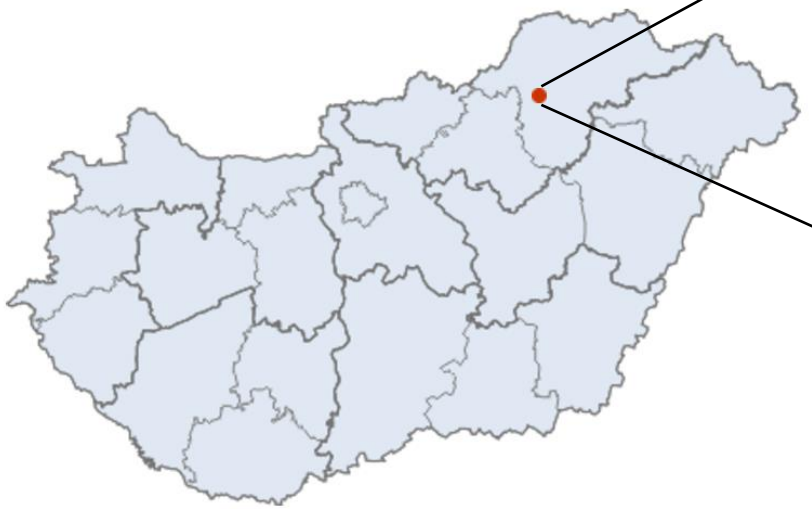


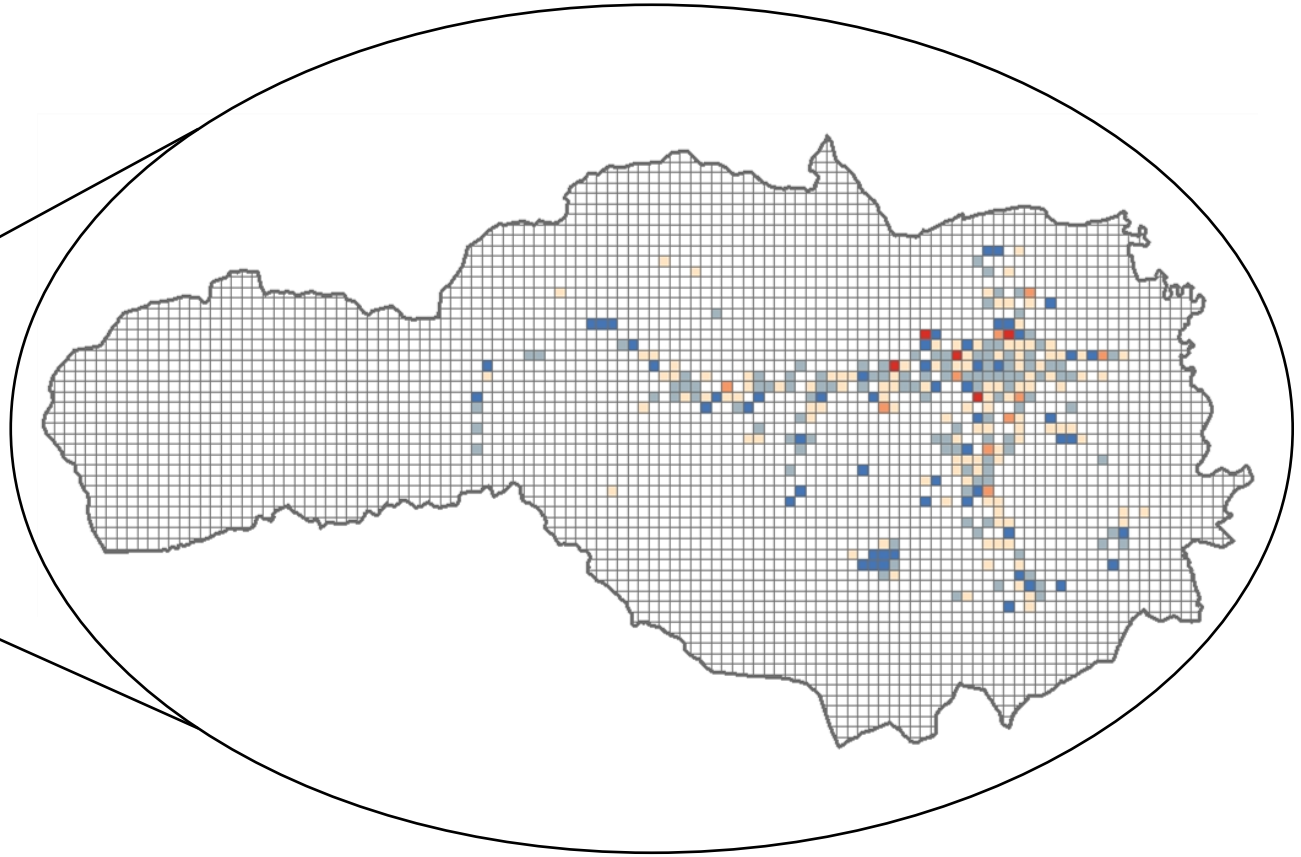
…для каждого отдельного населенного пункта , например: Озд

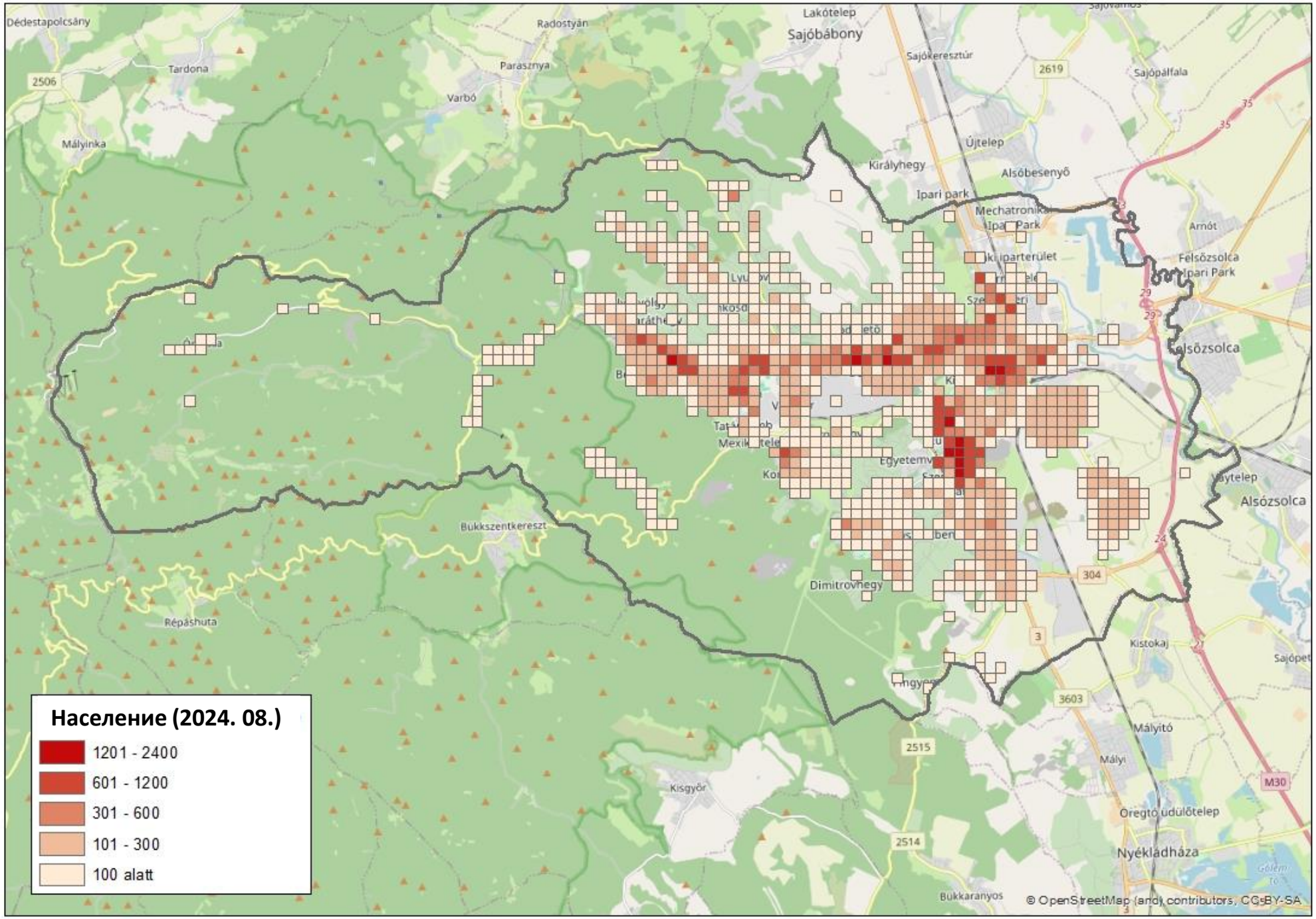


Мишкольц в деталях

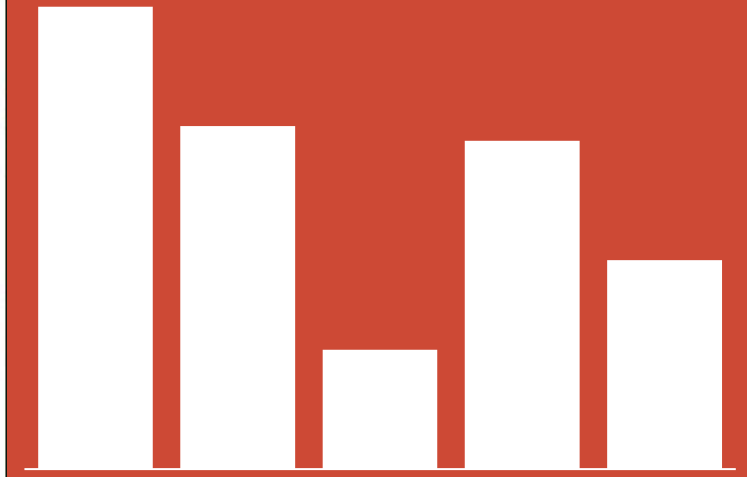
- Региональный центр
- (~150 000 жителей)
- четвертый по величине город Венгрии.



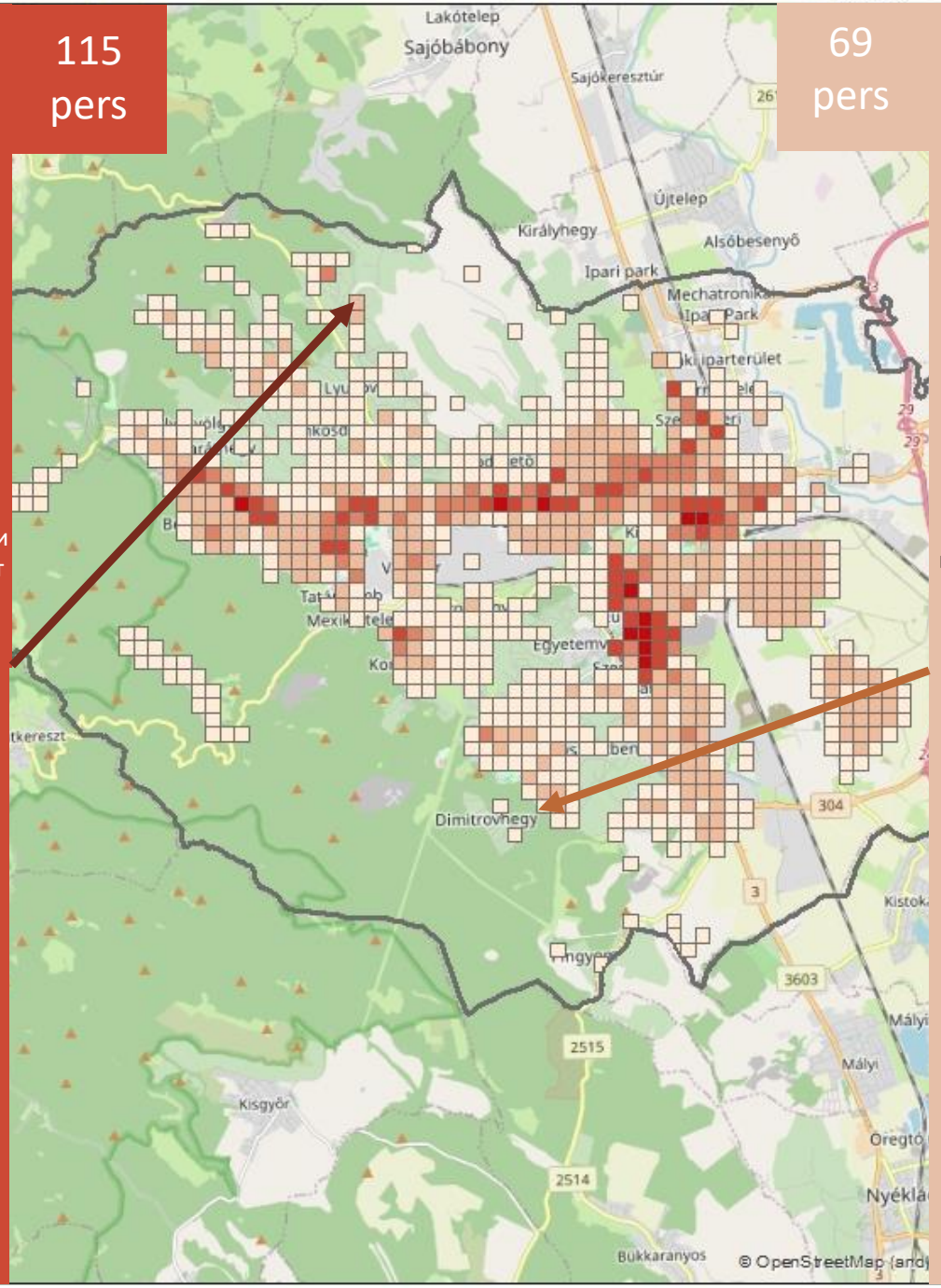




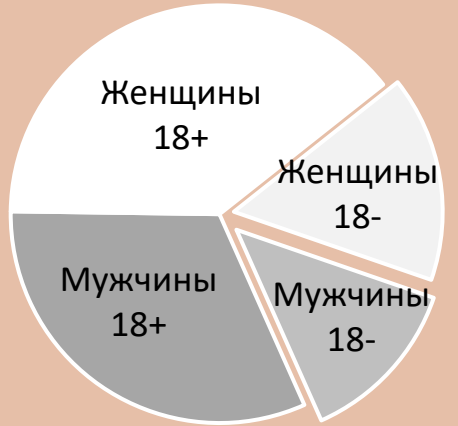
Плотность населения
(сетка 250x250 метров)



115 pers



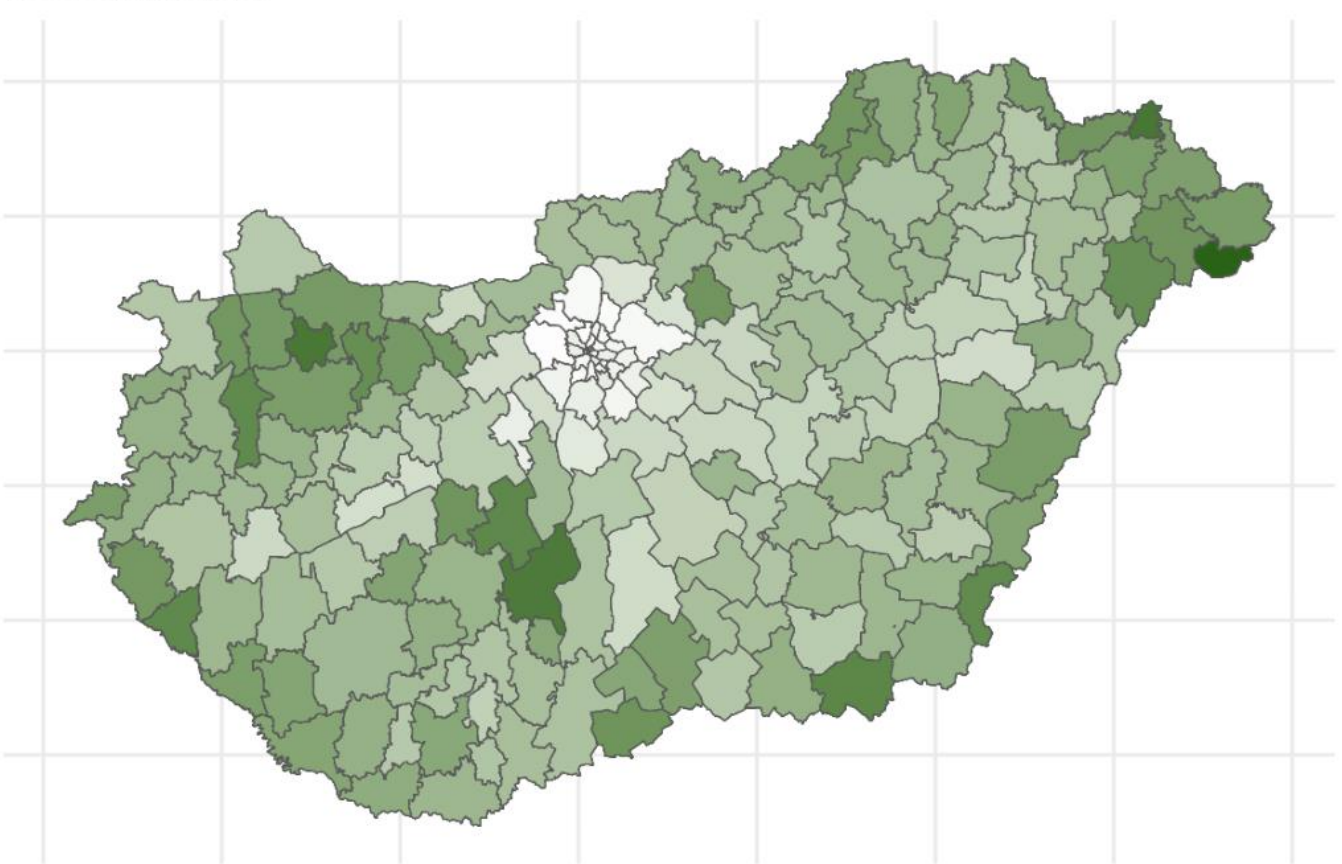
69 pers



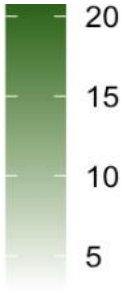
Региональная концентрация рынка труда

- **Источники данных:**
  - **Регистр учета населения (идентификационные и адресные данные) – Министерство внутренних дел**
  - **Налоговые данные по заработной плате – налоговые органы**
  - **Регистр предприятий (бизнес-регистр)**

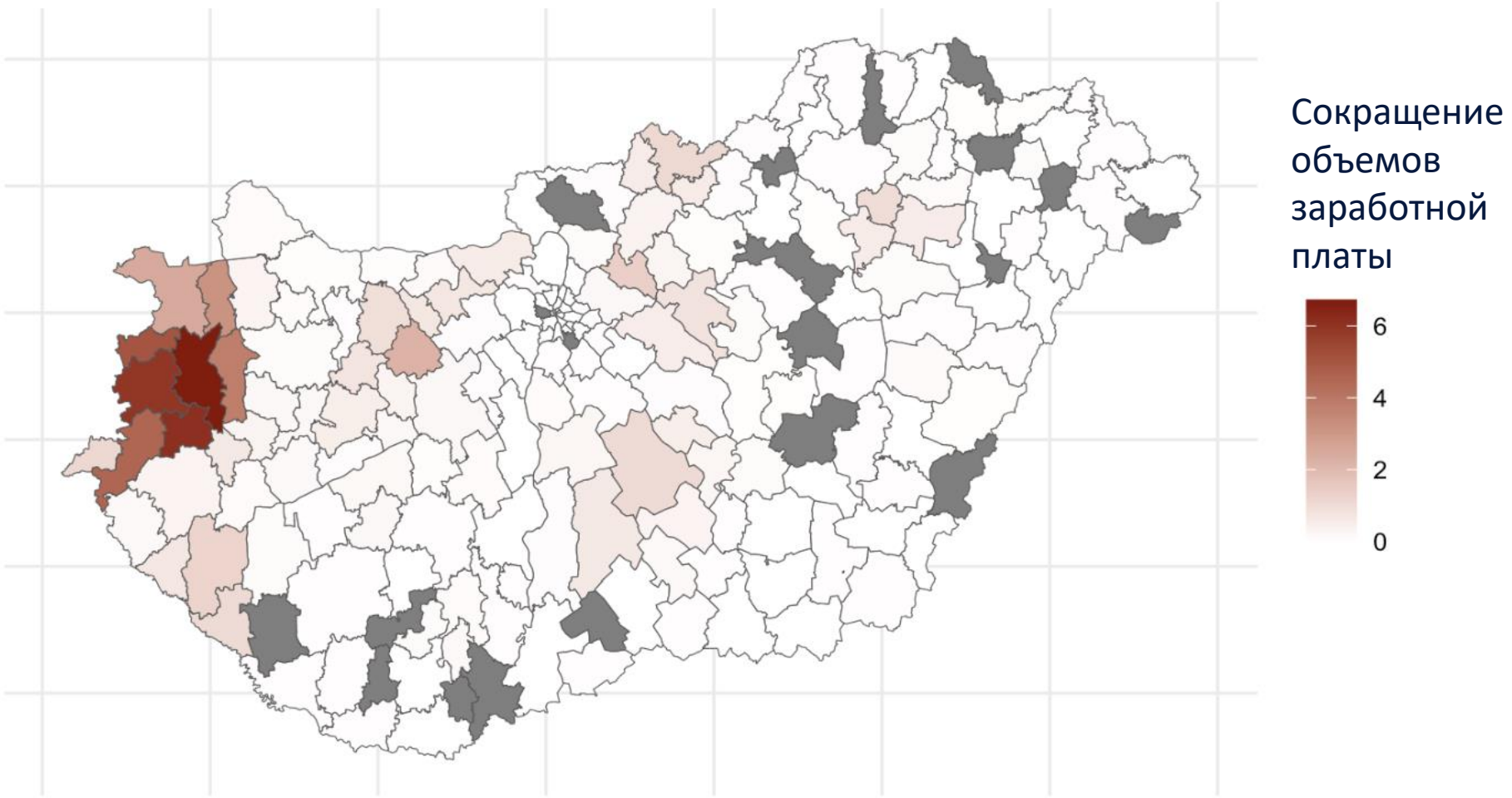
Концентрация районных рынков труда



Доля зависимости заработной платы от 10 крупнейших предприятий, %



Что произойдет, если компании, занимающиеся производством автомобилей (КДЕС 29), сократят численность своей рабочей силы на 10% из-за неблагоприятной ситуации в автомобильном секторе? В каких районах это больше всего повлияет на уровень заработной платы?



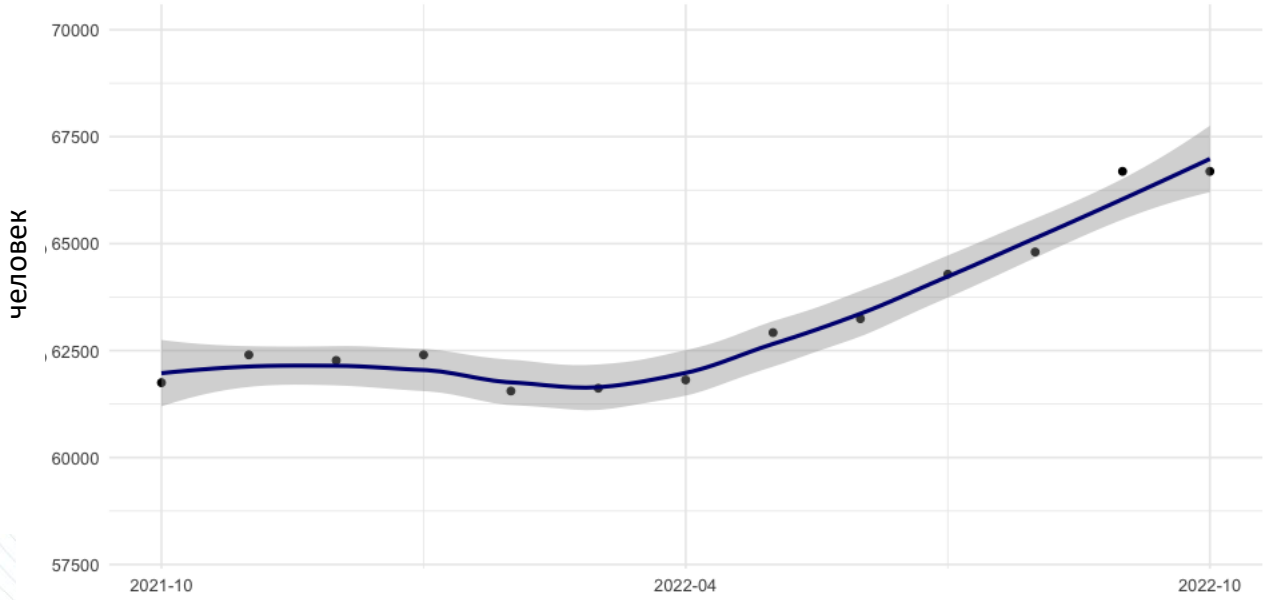
«Карта» занятости

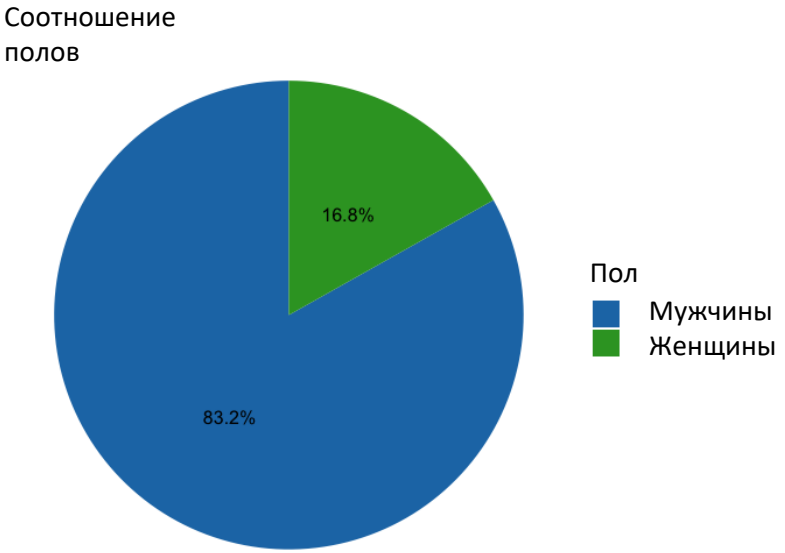
- **Используемые источники данных:**
  - **Регистр учета населения (идентификационные и адресные данные) – Министерство внутренних дел**
  - **Налоговые данные по заработной плате – Налоговые органы**
  - **Выплата пенсий – Государственное казначейство**
  - **Лица, ищущие работу – Министерство труда**

Аналитики, разработчики программного обеспечения и приложений

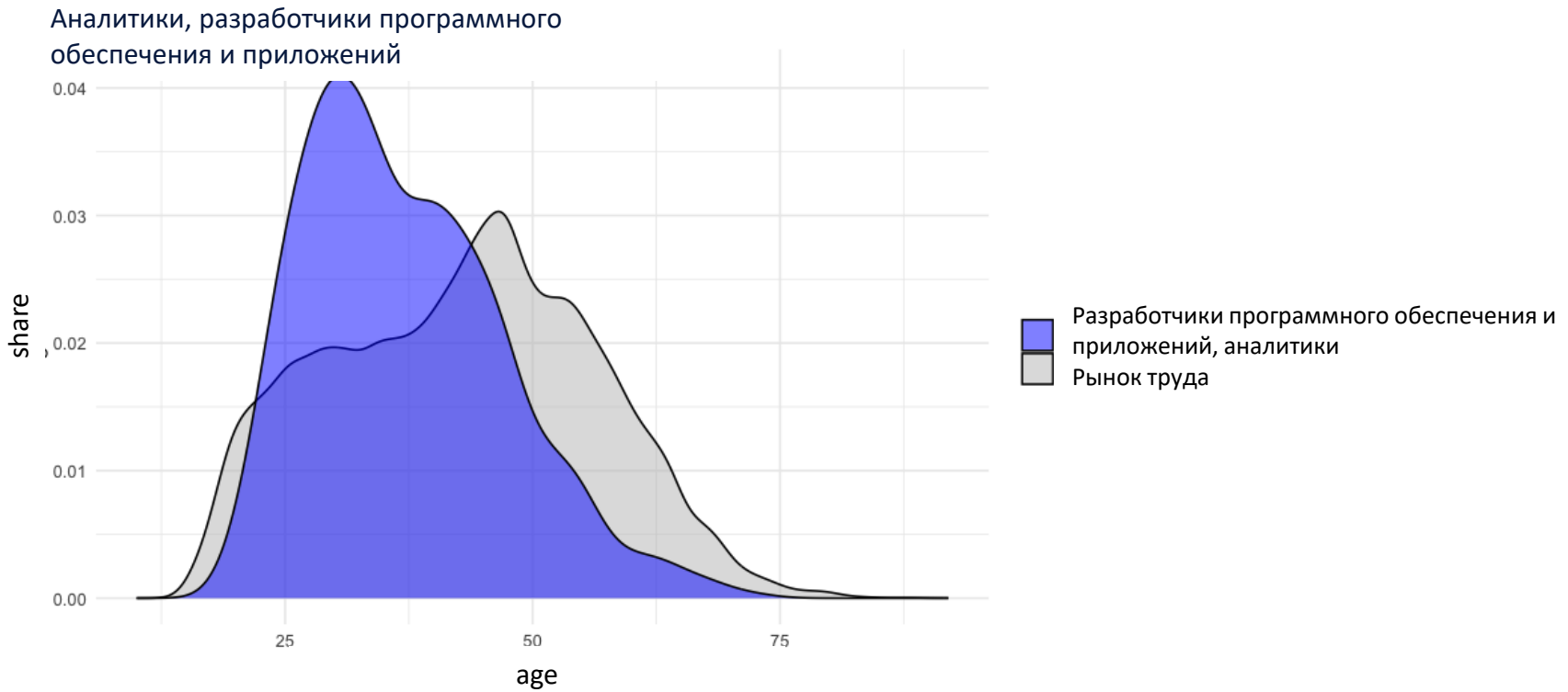
Численность занятых и соотношение полов

Аналитики, разработчики программного обеспечения и приложений



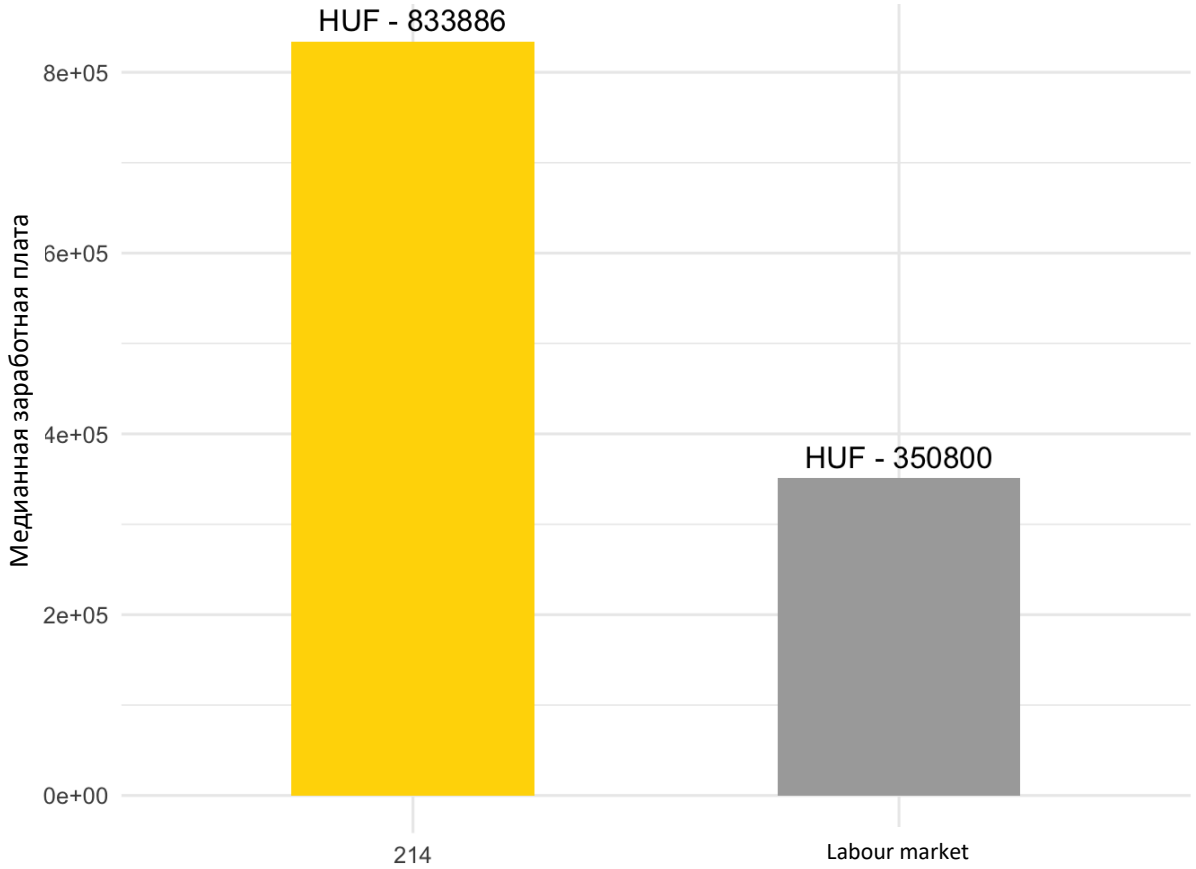


Распределение по возрасту

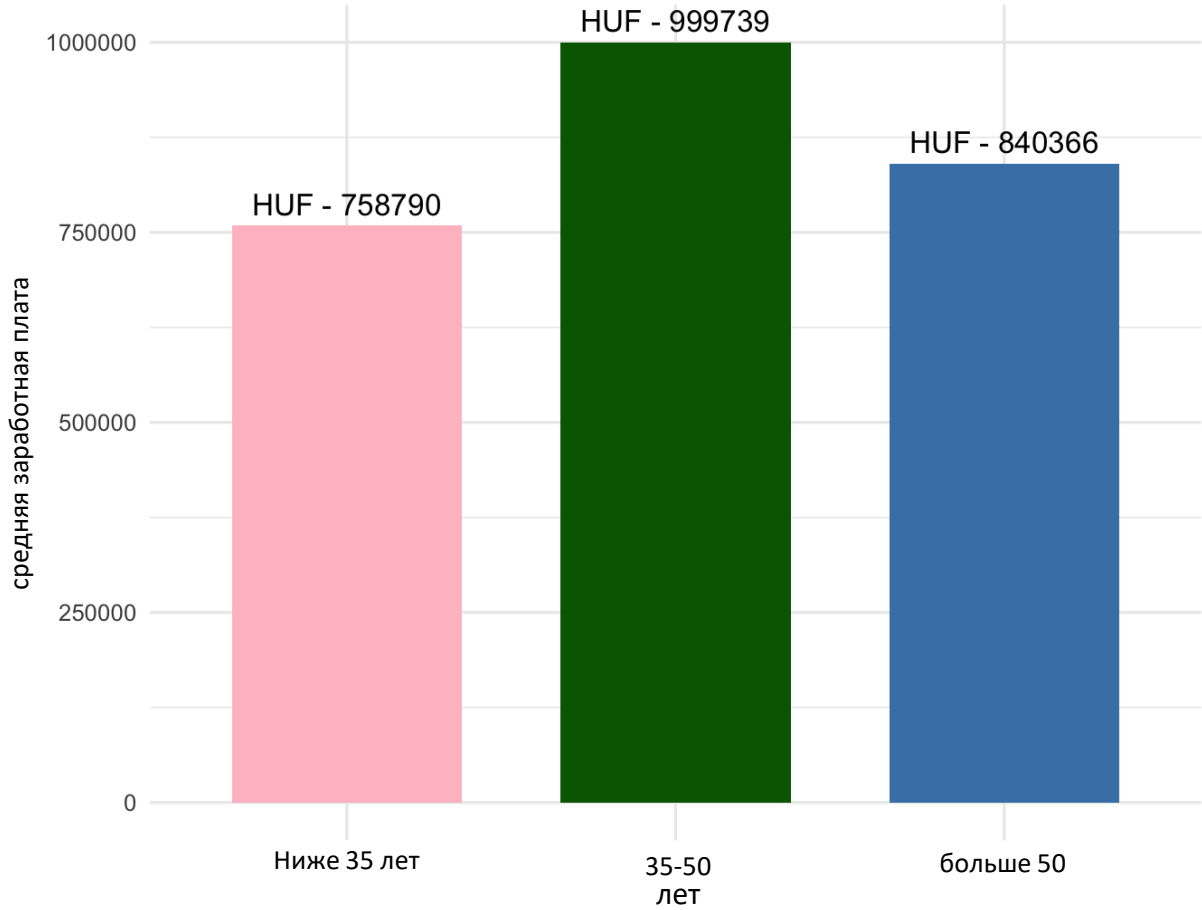


Заработная плата

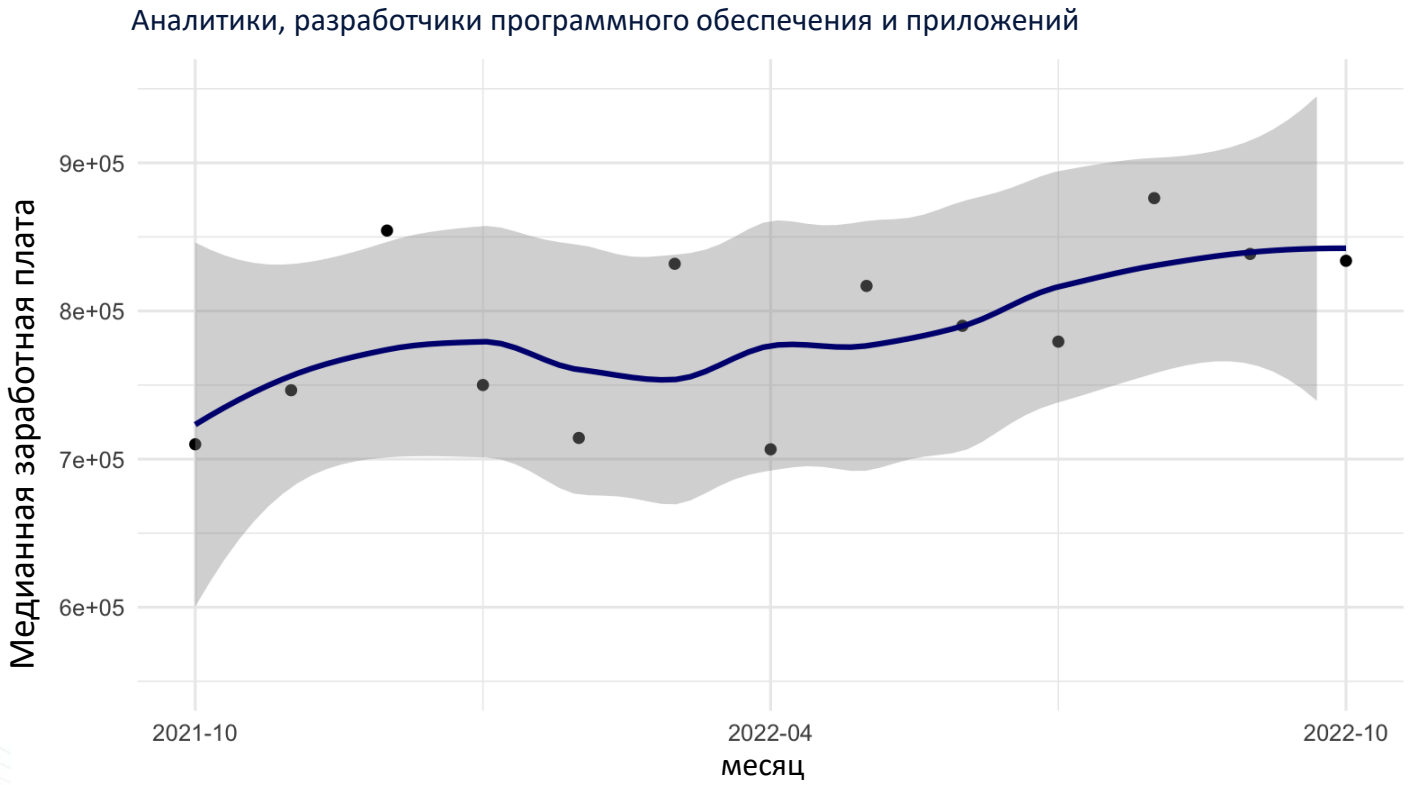
Аналитики, разработчики программного обеспечения и приложений



Аналитики, разработчики программного обеспечения и приложений



Динамика заработной платы и риск безработицы



Индикатор риска безработицы *

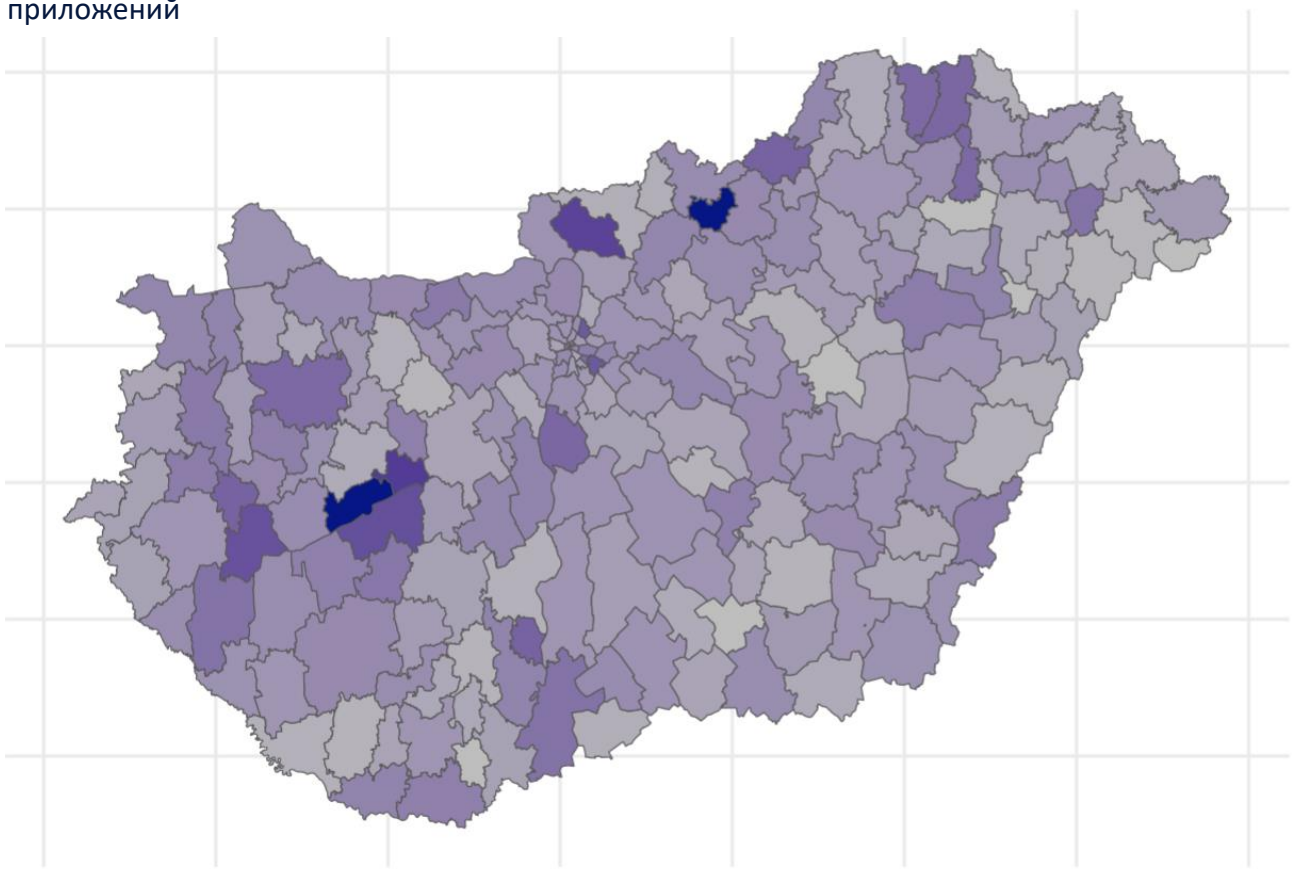
0.18

НИЗКИЙ

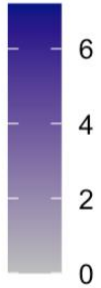
**Доля лиц, искавших работу (за 6 мес.), в общей численности ищущих работу.*

Территоральное распределение

Аналитики, разработчики программного обеспечения и приложений

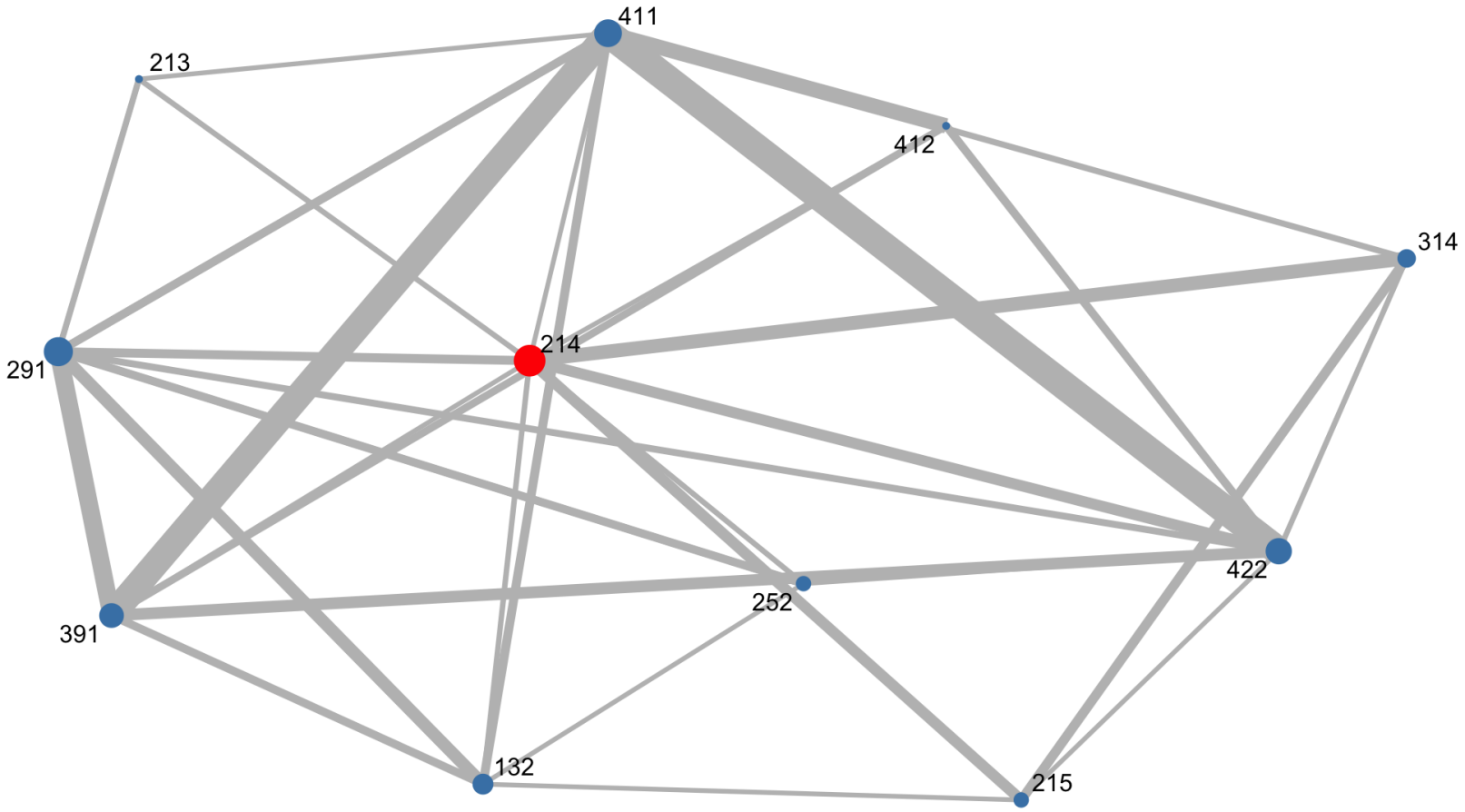


Концентрация рынка труда, %



Профессиональные переходы

Аналитики, разработчики программного обеспечения и приложений



- **214** **Аналитики, разработчики программного обеспечения и приложений**
- 132 Руководители подразделений сферы услуг
- 213 Прочие инженеры
- 215 Аналитики баз данных и сетей, операторы
- 252 Специалисты в области организационного управления и бизнес-политики
- 291 Другие высококвалифицированные администраторы
- 314 Специалисты в области информационных технологий и связи
- 391 Другие администраторы
- 411 Общие канцелярские и административные работники
- 412 Специалисты по бухгалтерскому учету
- 422 деятельность по работе с клиентами

Заключения

- Данные больше не являются дефицитом — они фрагментированы.
- Проблема заключается уже не в сборе данных, а в их интеграции.
- Искусственный интеллект преобразует данные в единую взаимосвязанную систему.
- Статистика переходит от статичных «снимков» к динамическому анализу.
- Интеграция микроданных и геопространственной информации открывает новый уровень понимания.
- От разрозненных показателей → к взаимосвязанным системам.

HUNGARIAN CENTRAL STATISTICAL OFFICE

Мы создаём не просто наборы данных — мы формируем систему.

Пал Бодай
pal.boday@ksh.hu
Центральное управление статистики Венгрии