Определение уязвимых групп населения

Согласно опросу «инФОМ» от декабря 2021 года, у 27 % россиян хватает денег только на еду, а ещё 9 % не могут позволить себе полноценное питание. Эти люди особенно внимательно следят за ценами, а темп роста цен на продукты обычно превышает средний темп инфляции. При этом Росстат считает, что расходы на продукты питания должны составлять примерно 36 % от среднемесячных расходов россиянина (ещё около 10 % приходится на услуги ЖКХ и жильё, 4 % — на лекарства).

До 2021 года «черта бедности» (жизнь на сумму ниже прожиточного минимума) в России определялась стоимостью минимальной продуктовой корзины. В том же году правительство «отвязало» уровень бедности от цен на базовые продукты: с 2021 года прожиточный минимум рассчитывается как 44.2 % от медианного дохода граждан РФ за прошлый год.

В вашем распоряжении есть данные о доходах, заболеваемости, социально незащищённых слоях населения России и другие экономические и демографические данные.

Ваша задача:

- кластеризовать регионы России и определить, какие из них наиболее остро нуждаются в помощи малообеспеченным/неблагополучным слоям населения;
- описать группы населения, сталкивающиеся с бедностью;
- определить:
  - влияет ли число детей, пенсионеров и других социально уязвимых групп на уровень бедности в регионе;
  - связаны ли уровень бедности/социального неблагополучия с производством и потреблением в регионе;
  - какие ещё зависимости можно наблюдать относительно социально незащищённых слоёв населения.

Задание «со звёздочкой»: если вы будете использовать данные из папки `crimes` (о ней — ниже), создайте из них полноценный датасет, приведя данные в удобочитаемый вид.

Данные

Все таблицы:

- `child_mortality_rural_1990_2021.xls` — число умерших на первом году жизни детей за год, по всем регионам, **в сельской местности**.
- `child_mortality_urban_1990_2021.xls` — число умерших на первом году жизни детей за год, по всем регионам, **в городской местности**.
- `disabled_total_by_age_2017_2022.csv` — число людей с инвалидностью по регионам, по месяцам, по возрастным группам.
- `morbidity_2005_2020_age_disease.xls` — заболеваемость на 100 тыс. человек населения, по возрастным группам и группам заболеваний.
- `poverty_percent_by_regions_1992_2020.csv` — процент людей, живущих за чертой бедности (с денежными доходами ниже величины прожиточного минимума), оценка за год по регионам.
- `welfare_expense_share_2015_2020` — расходы на социальную политику от общих расходов бюджета региона, % в год*.
- `cash_real_income_wages_2015_2020` — среднедушевые и реальные денежные доходы населения, номинальная и реальная начисленная зарплата, по регионам*.
- `poverty_socdem_20*.xls` — распределение малоимущего населения по социально-демографическим группам (дети, трудящиеся, пенсионеры) за 2017–2020 гг., по регионам.
- `housing_2020` — характеристика жилищных условий домохозяйств. Оценка домохозяйствами состояния занимаемого ими жилого помещения, обследование 2020 года*.
- `population.xlsx` — численность населения по регионам и федеральным округам на 1 января каждого года за 1999–2022 гг.
- `gross_regional_product_1996_2020.xls` — валовой региональный продукт на душу населения, в рублях.

- ➔ `regional_production_*_*.csv` — объём отгруженных товаров собственного производства или работ/услуг, выполненных собственными силами, по видам деятельности за 2005–2016 гг., 2017–2020 гг. (в тысячах рублей, значение показателя за год, полный круг).
- ➔ `retail_turnover_per_capita_2000_2021.xls` — оборот розничной торговли на душу населения, в рублях.

Папка crimes — сведения о преступлениях, совершённых отдельными категориями лиц за 2016–2022 гг., по месяцам, регионам, категориям лиц, категориям преступлений:

1. В папке содержатся файлы о расследованных преступлениях (число преступлений).
2. Файлы представлены в том виде, в котором их предоставляет Генпрокуратура. С ними требуется поработать, чтобы привести их в пригодный для анализа вид.
   Пример представления:

ФОРМА 4-ЕГС

Раздел 4. Сведения о преступлениях, совершенных отдельными категориями лиц (по расследова

Строка 1: Всего

Период: январь - июнь 2016 года

	Количество предварительно расследованных преступлений в отчетный период (из числа находившихся в произв					
	несовершеннолетними или при их соучастии	ранее совершавшими преступления	в том числе	группой лиц	группой лиц по предварительному сговору	организованной группой
			ранее судимыми			
	1	2	3	4	5	6
Российская Федерация	27 251	369 473	200 490	3 596	44 261	7 385
Центральный федеральный округ	4 025	69 777	40 131	801	10 453	3 311
Белгородская область	149	2 546	1 425	20	293	110
Брянская область	205	3 161	1 661	20	335	70
Владимирская область	319	3 533	2 074	68	603	77
Воронежская область	327	4 607	2 164	14	566	32
Ивановская область	190	2 432	1 230	17	316	91
Калужская область	199	2 045	1 311	28	491	46
Костромская область	133	2 144	1 114	33	277	61
Курская область	198	3 192	1 698	26	323	9
г. Москва	402	11 963	7 853	70	2 664	2 367

3. Данные за каждый месяц агрегированы с предыдущими месяцами этого года (то есть январь 2022, далее — январь + февраль 2022 и так далее).

4. Файлы содержат указание на месяц в названии файла по типу `4-EGS_Razdel_%_%MMYYYY%.xls` (в названии может быть раздел 4 или 5). Это связано с тем, что форма статистического наблюдения изменилась с 2021 года, поэтому файлы различаются, однако по большому срезу данных в них представлена схожая информация.

   Категории лиц, по которым Генпрокуратура представляет статистику:

   a. Несовершеннолетние или преступления, совершаемые при их соучастии.
   b. Ранее совершавшие преступления.
      i. Из них ранее судимые.
   c. Группа лиц.
   d. Группа лиц по предварительному сговору.
   e. Организованная группа.
   f. Преступное сообщество (преступная организация).
   g. В состоянии опьянения.
   h. Член семьи, супруг, сожитель, сексуальный партнёр (в том числе бывший).
      i. Из них в отношении женщин.
      ii. Из них в отношении несовершеннолетних.
   i. Иностранные граждане и лица без гражданства.
      i. Из них граждан государств-участников СНГ.
      ii. Из них трудовых мигрантов.
      iii. Из них незаконных мигрантов.

Обратите внимание, что пункты *c-f* не суммируются в методологии Генпрокуратуры. Однако при необходимости вы можете их суммировать, введя новую категорию.

Обратите внимание, что пункты *h-i* появляются в статистике только с начала 2021 года.

- Вся информация в сумме приводится на первом листе каждого файла («строка 1» в терминологии документов). На остальных строках в зависимости от года приводится разбивка по категориям преступлений (особо тяжкие, тяжкие, средней тяжести, небольшой тяжести), а далее, если есть дополнительные листы (с 2021 года) — по главам и статьям Уголовного кодекса. Для более ранних годов разбивки по преступлениям нет — только по категориям преступлений.
- Справка об уголовной ответственности несовершеннолетних. Следует учитывать, что с разного возраста наступает подсудность по разным статьям. Для одних статей ответственность наступает с 14 лет, для других — с 16 лет.

`drug_alco` — сведения о заболеваемости алкоголизмом и наркоманией, на 100 тыс. населения (2005–2018):

- Речь идёт о впервые установленном диагнозе, в данном случае — о постановке на учёт в диспансере. То есть если в прошлом году человек был признан алкоголиком, в этом году он уже не учитывается в статистике, если стоит на учёте или вылечился и был снят с учёта. Также если человек не попал на учёт, хотя соответствует критериям постановки соответствующего диагноза, он не будет учтён в данной статистике. Таким образом, эти цифры занижены. Кроме того, так как показатель приводится именно по впервые выявленным случаям, в теории его можно аккумулировать (но с оговорками).
- Показатель на 100 тысяч населения — привычный показатель в эпидемиологии. Однако очевидно, что в данном случае он рассчитывается по всей популяции, хотя, возможно, такой диагноз ставится чаще в некоторых когортах.
- В файле четыре листа. Отдельно приводятся показатели за 2017–2018 гг., так как взяты из другого источника, и там немного иначе написаны названия регионов.

`newborn_2006_2022_monthly.csv` — рождённые в этом месяце, по регионам, без учёта мертворождённых.

* Файлы *Google Spreadsheets*. В каждом файле — по несколько листов, описание данных на листах и источники данных — на листе `desc` в каждом файле.

Ссылки на источники данных:

- Число умерших на первом году жизни детей за год.
- Численность населения с доходами ниже величины прожиточного минимума, в процентах от общей численности населения.
- Численность инвалидов по возрастным группам в разрезе субъектов РФ.
- Распределение малоимущего населения по социально-демографическим группам.
- Заболеваемость с диагнозом, установленным впервые в жизни, проживающих в районе обслуживания лечебного учреждения, на 100 тыс. человек населения.
- Комплексное наблюдение условий жизни населения в 2020 году.
- Социальное положение и уровень жизни населения России в 2021 году.
- Численность постоянного населения на 1 января.
- Оборот розничной торговли на душу населения с 2000 г.
- Статистика преступности в России, открытые данные Генпрокуратуры.
- Число зарегистрированных родившихся.

Дополнительные ресурсы:

- Портал «Если быть точным» — данные и аналитика о социальных проблемах.
- Росстат — государственное статистическое ведомство.
- Портал статистических данных РФ.

Обратите внимание:

1. Методологии подсчёта некоторых показателей могли меняться за время их наблюдения.
2. Не всегда один и тот же регион пишется одинаково в разных датасетах. Такое очень часто происходит с написаниями регионов РФ в разных документах, а также в других странах (их нет в данной задаче). Это необходимо учитывать при объединении датасетов.
3. Особенностью Тюменской области является то, что в её состав входят два других субъекта Российской Федерации: Ханты-Мансийский автономный округ — Югра и Ямало-Ненецкий автономный округ. При этом чаще всего в датасетах указывается отдельно Тюменская область и отдельно — без учёта автономных округов.
4. То же самое касается Архангельской области — в её состав входит самостоятельный субъект Ненецкий автономный округ. Способы написания такие же, как с Тюменской областью.
5. Нежелательно использовать данные по федеральному округу в совокупности. Во-первых, это агрегированные данные, и чаще всего по ним нельзя сделать никаких интересных выводов, а во-вторых, на протяжении последних 20 лет составы федеральных округов менялись (некоторые регионы меняли принадлежность к округу), из-за чего в них мог произойти резкий рост/снижение численности населения, что не отображало какое-то демографическое явление.