

# Анализ структуры доходов граждан в разрезе муниципальных образований в 2018-2022 годах.

Ярослав Снарский

31 10 2024

## Работа с датасетами и предобработка данных

На первом этапе я загружаю необходимые данные - датасет об объеме социальных выплат и налогооблагаемых доходов (далее - датасет ростата) и версионный справочник муниципальных образований (далее - версионный справочник). Замечу, что скаченный по ссылке в ТЗ датасет с социальными выплатами и доходами не содержит 2023 год, поэтому в качестве финального года я буду брать 2022 в последующих заданиях. Полный код для предобработки доступен в приложенном Rmd файле.

Я собираюсь объединить два датасета на основе кода ОКТМО, т.е. каждой версии ОКТМО из справочника, соответствующей определенному году, я буду сопоставлять ОКТМО из датасета ростата за определенный год. Такой подход позволяет быстро покрыть около 95% всех муниципалитетов в каждом году. Оставшиеся муниципалитеты я на данном этапе не рассматриваю, потому что предполагаю, что их можно считать случайными пропусками. В будущем можно настроить эту систему более хитро, - например, после объединения датасетов по ОКТМО несовпадения можно попытаться сопоставить на основе названия и региона. Я предварительно пробовал этот вариант - даже после повторного мэтча по названию остается около 5% несопоставленных муниципалитетов. Общая идея этого подхода - после сопоставления по ОКТМО в каждом году для большинства муниципалитетов мы получаем новую колонку из версионного справочника - **territory\_id**. По ней можно будет отследить непрерывный ряд, поскольку это уникальный индикатор территории, который неизменен при изменении типа, названия и кода, но меняется при изменении границ - это те условия, которые требуются в ТЗ.

Я предварительно не предобрабатывал датасеты в Excel, поэтому переложим эту ответственность на R. Напишем функцию, которая учитывает все возможные формы написания ОКТМО в датасете ростата.

```
# Function to standardize ID format
standardize_oktmo <- function(id) {
  # handling missing values
  if (is.na(id)) {
    return(NA)
  }

  # removing any existing non-digit characters
  id <- gsub("[^0-9]", "", id)

  # checking if the ID has length 2 (e.g., "01", "99")
  if (nchar(id) == 2) {
    return(id) # Keep as is
  }

  # checking if the ID has length 7 (for cases like "3601000")
  if (nchar(id) == 7) {
    # adding a leading "0" to make it 8 characters
    id <- paste("0", id, sep = "")
  }

  # checking if the ID has length 8
  if (nchar(id) != 8) {
    stop("ID must be 8 characters long")
  }

  # standardizing the ID
  id <- sprintf("%08d", id)
}
```

```

id <- paste0("0", id)
}

# splitting the ID into the required format: "xx-xxx-xxx-xxx" by inserting "-"
formatted_id <- paste0(substr(id, 1, 2), "-", substr(id, 3, 5), "-", substr(id, 6, 8), "-000")
return(formatted_id)
}

```

Дальше я получаю из данных ростата 2018 года список всех регионов, чтобы присвоить каждому муниципалитету его региональную привязку для дальнейшего анализа, в т.ч. обрабатывая случаи НАО, ХМАО, ЯНАО. Весь код для препроцессинга - в Rmd файле.

```

#function to process each dataframe
process_data <- function(df, oktmo_dict) {
  # standardising oktmo and creating oktmo_id
  df$mun_name <- gsub(" - всего", "", df$mun_name)
  df$mun_name <- trimws(df$mun_name, which = "r")
  df$oktmo <- ifelse(df$mun_name == "Ненецкий автономный округ", "11", df$oktmo)
  df$oktmo_mod <- ifelse(df$mun_name == "Ненецкий автономный округ", "118",
    ifelse(df$mun_name == "Ханты-Мансийский автономный округ", "718",
      ifelse(df$mun_name == "Ямало-Ненецкий автономный округ", "719", df$oktmo)))
  df$oktmo_id <- sapply(df$oktmo, standardize_oktmo)
  # assiging region based on oktmo_dict
  # and conditional expression to assign specific cases:
  df$region <- ifelse(
    substr(df$oktmo_mod, 1, 3) == "118", "Ненецкий автономный округ",
    ifelse(substr(df$oktmo_mod, 1, 3) == "718", "Ханты-Мансийский автономный округ",
      ifelse(substr(df$oktmo_mod, 1, 3) == "719", "Ямало-Ненецкий автономный округ",
        ifelse(nchar(df$oktmo_mod) == 3,
          oktmo_dict[substr(df$oktmo_mod, 1, 3)],
          oktmo_dict[substr(df$oktmo_id, 1, 2)])))
  return(df)
}

```

После загрузки справочника Сбериндекса отсортируем наблюдения, согласно документации.

```

sberindex_data_2018 <- sberindex_data[sberindex_data$year_from <= 2018 & sberindex_data$year_to > 2018, ]
sberindex_data_2019 <- sberindex_data[sberindex_data$year_from <= 2019 & sberindex_data$year_to > 2019, ]
sberindex_data_2020 <- sberindex_data[sberindex_data$year_from <= 2020 & sberindex_data$year_to > 2020, ]
sberindex_data_2021 <- sberindex_data[sberindex_data$year_from <= 2021 & sberindex_data$year_to > 2021, ]
sberindex_data_2022 <- sberindex_data[sberindex_data$year_from <= 2022 & sberindex_data$year_to > 2022, ]

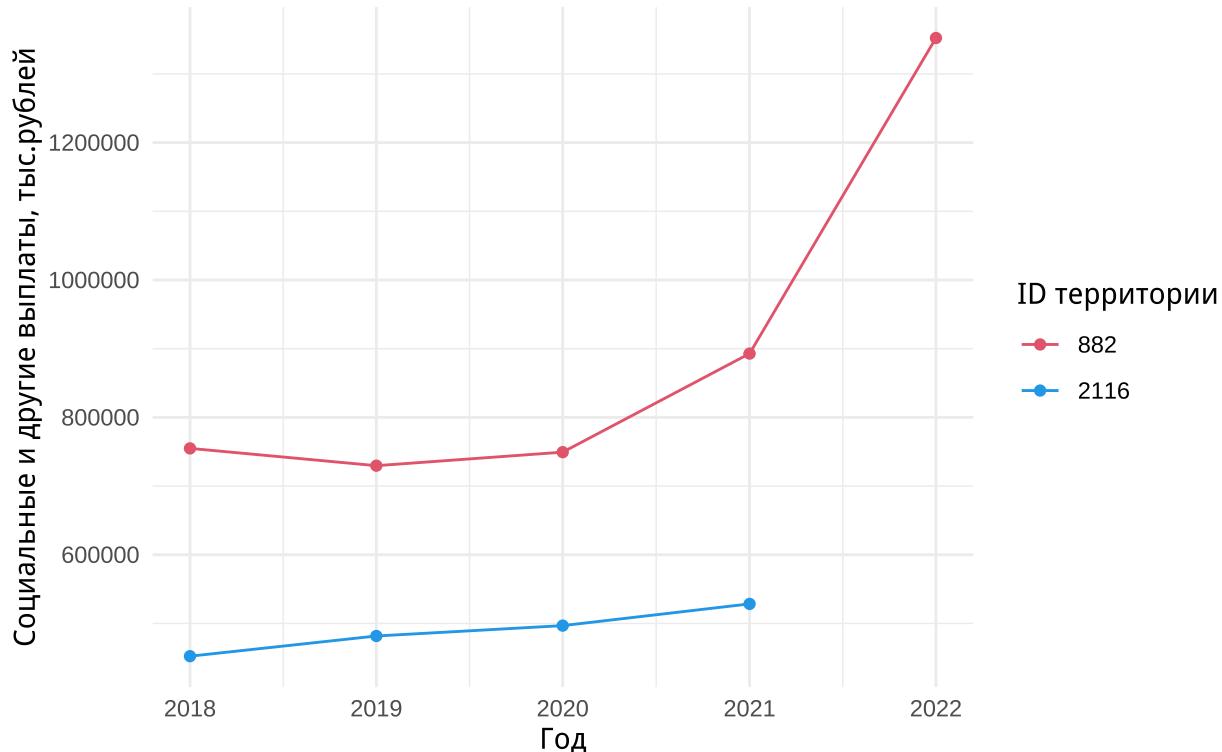
```

## Построение рядов на основе ID муниципалитета

Объединим по ОКТМО все муниципалитеты за определенный год из двух датасетов, следя логике версционности, - слияние по ОКТМО происходит в том случае, если муниципалитет в версионном справочнике входит во временные границы, задаваемые значениями колонок year\_from и year\_to.

oktmo	mun_name	territory_id	region	year	taxable_income	social_payments
57-605-000-000	Александровский	734	Пермский край	2018-01-01	2384197	2275206
57-605-000-000	Александровский	734	Пермский край	2019-01-01	2510037	2320961
57-502-000-000	Александровский	734	Пермский край	2020-01-01	2635059	2653125
57-502-000-000	Александровский	734	Пермский край	2021-01-01	2816591	2830495
57-502-000-000	Александровский	734	Пермский край	2022-01-01	3027614	3414275

## Динамика социальных выплат в муниципалитетах Жарковский (Тверская область) и Ромненский (Амурская область)



Получаем ряд 882 (10-540-000-000) выше, для которого выполняются требования: при условии, что название ОКТМО менялось, у территории одинаковый индекс **territory\_id** на протяжении всех пяти лет наблюдений. Однако такого результата удалось достичь не для всех муниципалитетов. В нескольких рассмотренных мною случаях проблема сводилась к тому, что в датасете ростата был старый ОКТМО, в то время как в версионном справочнике уже был прописан новый. Муниципалитет с ID 2116 (28-614-000-000) как раз обрывается в 2021 году, поскольку в 2022 в справочнике он имеет ОКТМО 28-614-000-000, а в датасете ростата 28-514-000-000, ввиду чего теряется часть информации. Потенциально проблему могло бы решить сопоставление как по ОКТМО, так и по названию (с контролем на регион, чтобы не наткнуться на дублирующие названия в разных регионах), но не всегда название в версионном справочнике совпадало с названием в датасете ростата. Например, *Нерехтский и г.Нерехта* (ростат) vs. *Нерехта и Нерехтский* (справочник). Такие ограничения не позволяют на сто процентов сопоставить все муниципалитеты.

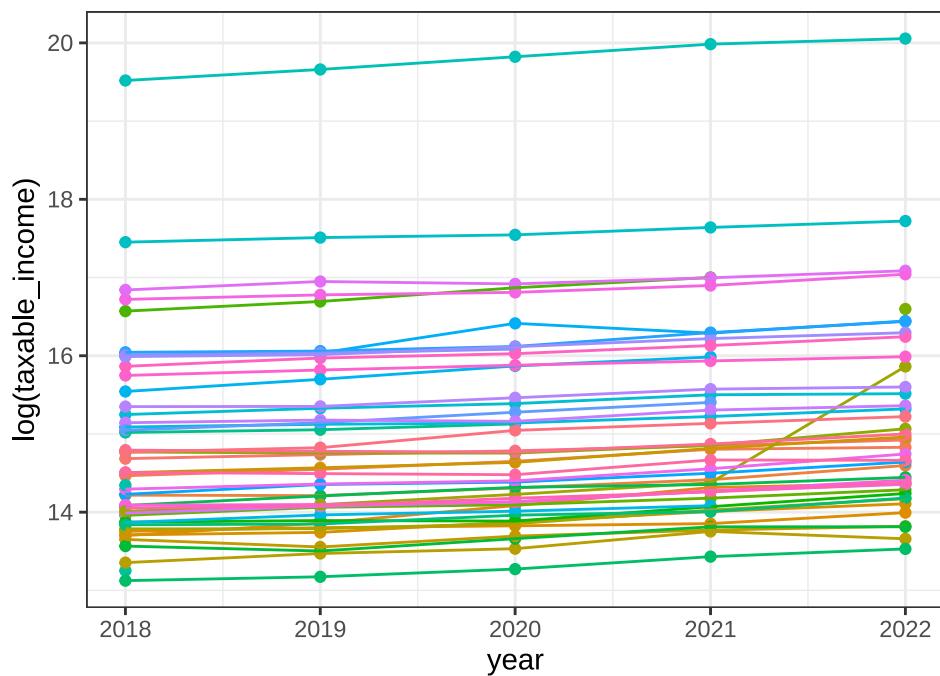
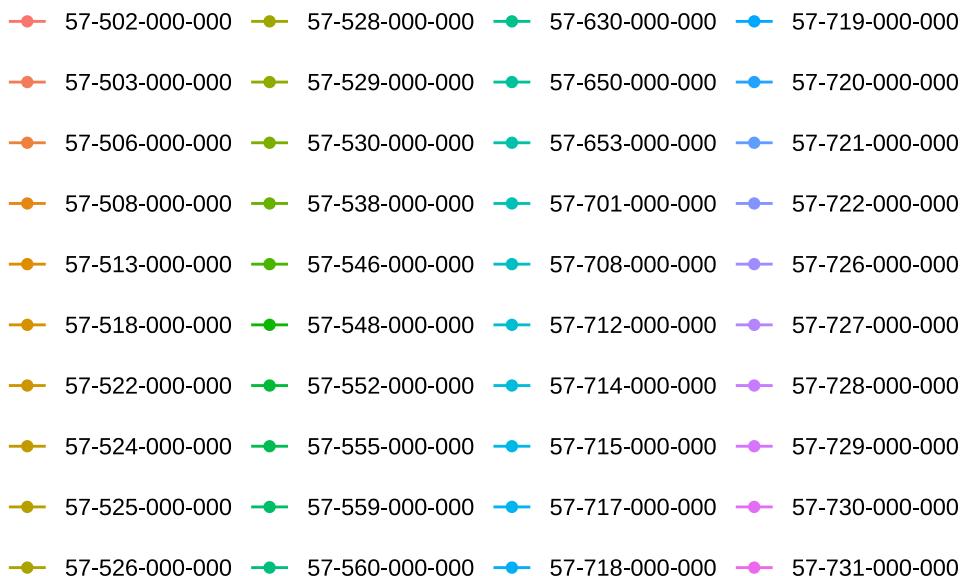
Если я правильно понял, то под присвоением актуального действующего кода ОКТМО подразумевается, что мы должны назначить самый поздний ОКТМО для **territory\_id**. Ниже я предлагаю код, который это делает, и визуализацию результатов. Теперь даже в случае изменения ОКТМО или названия муниципалитета, он имеет постоянное наименование по самому недавнему коду ОКТМО.

```

# Отфильтруем справочник, чтобы оставить только самую позднюю запись для каждого territory_id
latest_reference <- sberindex_data %>%
  filter(!is.na(territory_id)) %>%
  group_by(territory_id) %>%
  filter(year_from == max(year_from)) %>%
  ungroup() %>% select(oktmo, territory_id) %>% rename(oktmo_new = oktmo)

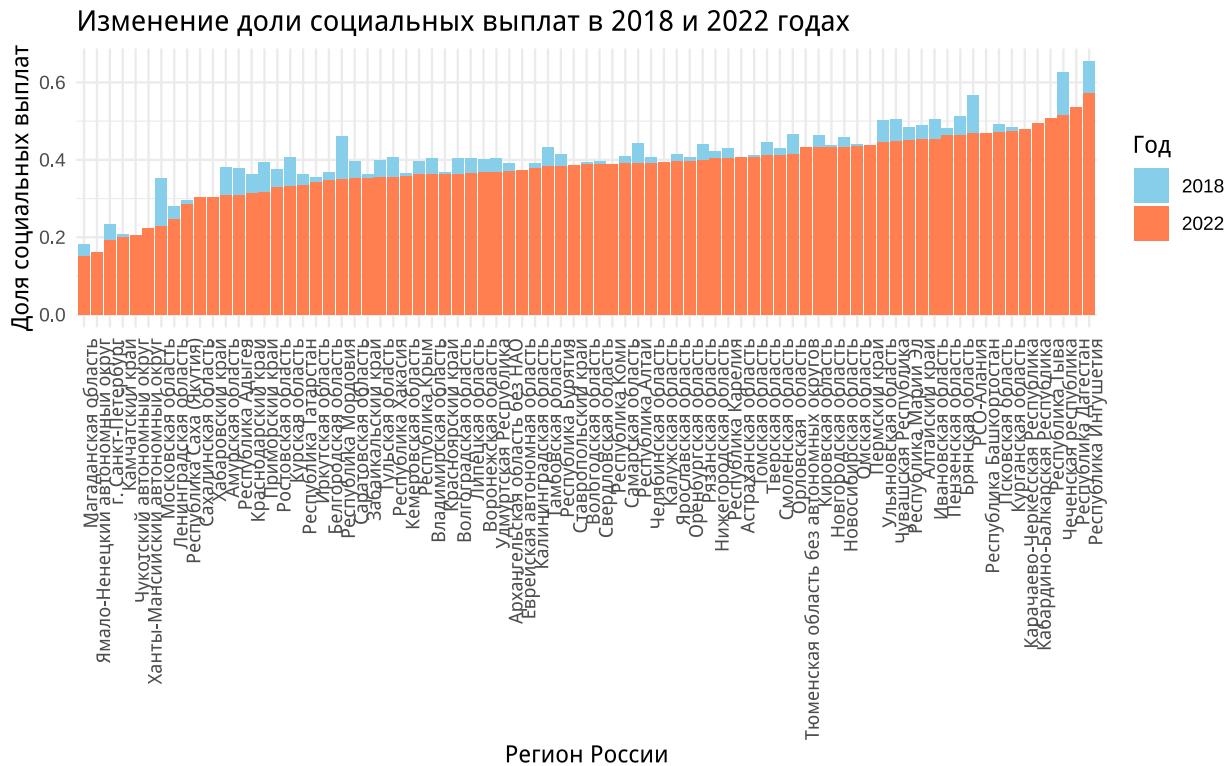
# Соединяем основной датасет с отфильтрованным справочником по territory_id и year
final_data <- df_stats_merged %>% filter(!is.na(territory_id)) %>%
  left_join(latest_reference, by = "territory_id")

```

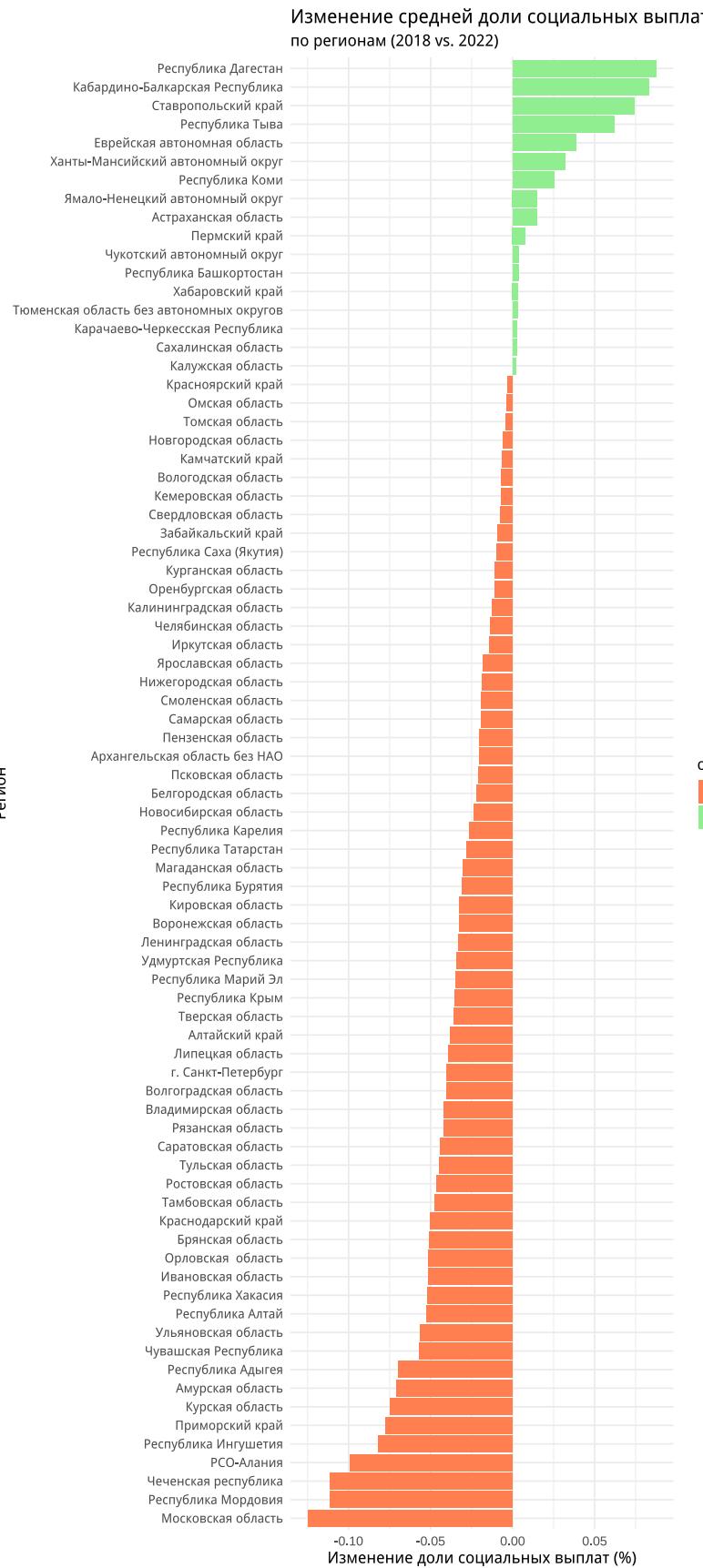


## Дескриптивный анализ и визуализация

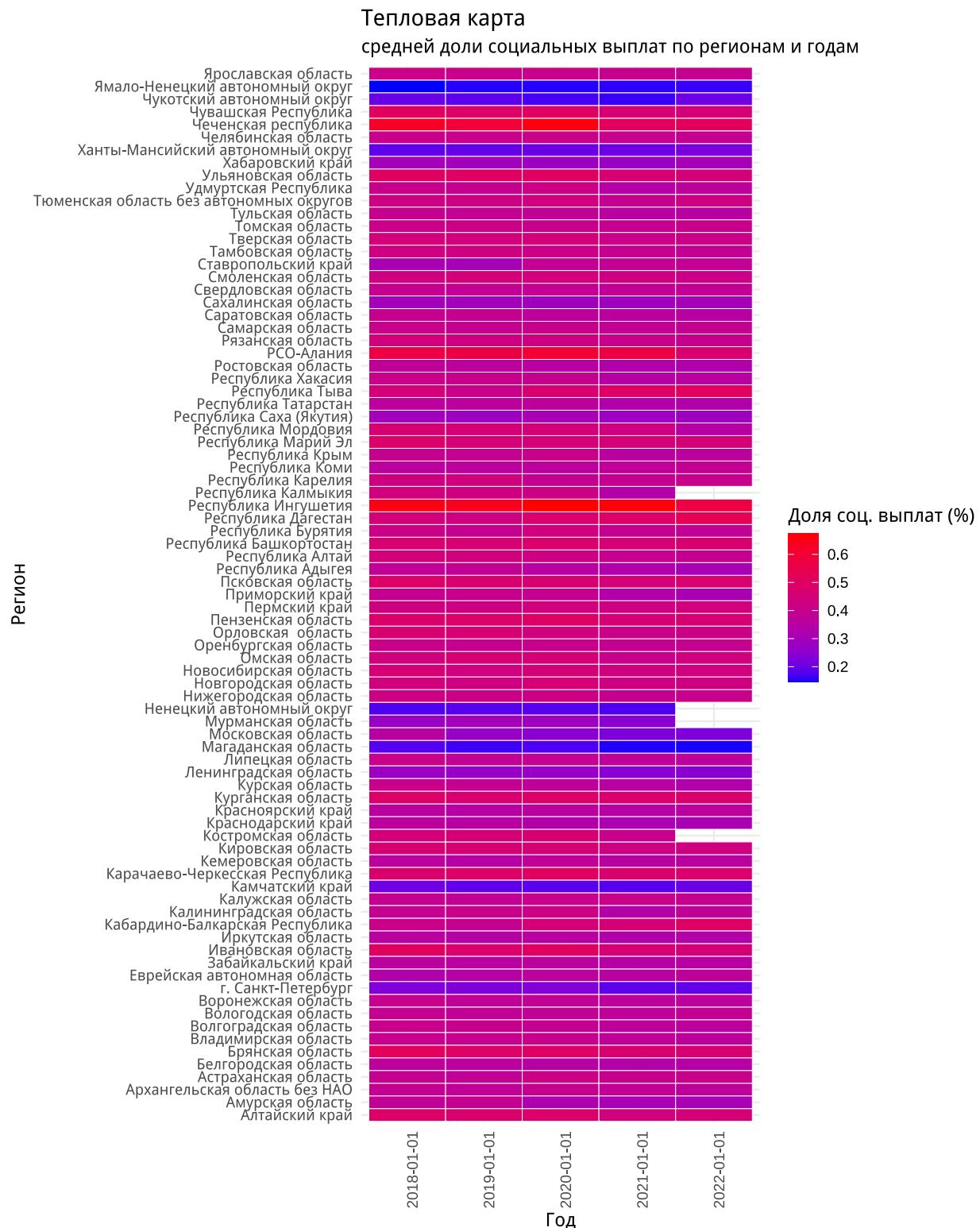
Столбчатая диаграмма ниже визуализирует изменение в доли социальных выплат в 2018 и 2022 годах. Общий тренд - на снижение этой доли в 2022 году по сравнению с 2018 г. Такое снижение может диктоваться двумя механизмами - либо люди стали больше зарабатывать (увеличилась доля других частных доходов), либо уменьшились социальные расходы государства. Уменьшение социальных расходов может быть связано, например, с уменьшением уровня безработицы, снижением рождаемости (меньше семей соответствуют условиям выплат маткапитала), уменьшением доли тех, кому положены пенсии (Пенсионная реформа 2019 года и повышенная смертность от COVID-19), переходом от монетарной поддержки к материальной (выдача инвалидного кресла вместо эквивалентной денежной суммы). В целом, в российской социальной политике заметен тренд на адресную социальную поддержку, которая сужает охват целевых групп. Не исключено, что оба фактора - и увеличение частных доходов, и снижение социальных расходов государства - могли иметь место одновременно.



Чтобы отыскать, какие именно регионы все же увеличили долю социальных выплат, построим столбчатую диаграмму, которая отражает изменение средней доли социальных выплат по регионам. Оказывается, что всего 17 регионов - Дагестан, Кабардино-Балкария, Ставропольский Край, Тыва, среди прочего, - слегка увеличили долю выплат.



Наконец, чтобы проследить динамику во весь период с 2018 по 2022, я использую тепловую карту.



Эта карта показывает, что ростат по какой-то причине перестал давать статистику по ряду регионов (79 в 2022 vs 83 в 2018). Тепловая карта показывает, что внутри одного региона изменения в доли социальных выплат не

такие большие, как между регионами, что, с одной стороны, может говорить об институциональной инерции. С другой стороны, во многих регионах видно “посинение” - снижение доли социальных выплат в объеме доходов населения муниципальных образований.

## Гипотетический дизайн исследования и его проблемы

Дизайн исследования, направленного на оценку влияния инвестиционных проектов на налогооблагаемые денежные доходы граждан, предполагает несколько стратегических выборов на этапе планирования исследования. Я кратко опишу несколько из них: например, нужно «на берегу» определиться с тем, какого типа инвестиционные проекты попадут в исследование, поскольку нужно сузить понимание того, эффект каких именно инвестиционных проектов мы собираемся исследовать. Альтернативой могло бы быть включение типа инвестиционного проекта (например, проекты в сфере жилищного строительства, проекты транспортной инфраструктуры, экологические проекты и т.д.) в анализ, а именно предположение того, что эффект может быть гетерогенным в зависимости от типа проекта и механизма, через который происходит причинно-следственная связь. Если у нас есть информация по типам проектов, то мы могли бы оценить «средний» эффект реализации инвестиционного проекта, а затем посмотреть на то, какие эффекты производят инвестиционные проекты разного типа.

Это предварительное замечание связано с тем, что на этапе планирования исследования нужно выбрать теоретический и эмпирический эстиманды – те контрфактуальные величины, оценку которых мы бы хотели получить с помощью выбранного нами подхода. В качестве точки отсчета можно взять оценку среднего эффекта воздействия. Если в наиболее простом случае инвестиционные проекты реализовывались в муниципалитетах случайно, т.е. не зависели от прочих характеристик муниципалитетов, то простая разница средних давала бы несмещенную оценку. Однако такой взгляд не совсем реалистичен, поскольку такие проекты предполагают изучение местных (социальных, политических и бизнес-) условий инвесторами, а потому могут быть сильно завязаны на социально-экономические показатели муниципалитетов. Ввиду этой особенности при оценке эффекта простой разницей средних возникает так называемая ошибка отбора (*selection bias*). Так, способ отбора муниципалитетов для участия в программе будет определять дизайн исследования, поскольку если каждый муниципалитет с равной вероятностью может попасть в группу воздействия, то такие условия можно считать экспериментальными, а создавшуюся в результате вариацию в зависимой переменной атрибутировать именно воздействию (факту наличия инвестиционного проекта). Если же мы предполагаем, что есть внешние факторы, которые определяли шансы попадания муниципалитета в группу воздействия и они замерены, мы можем использовать техники мэтчинга, чтобы добиться сопоставимости группы воздействия и группы контроля по набору ковариат, замеренных до воздействия.

Другой вариант потенциального смещения оценки – если программа реализовывалась централизованно и разработчиками учитывались характеристики муниципальных образований при отборе, то муниципалитеты с определенными характеристиками с большей вероятностью могли получить воздействие (например, программа должна была «поднять» отстающие муниципалитеты до среднего уровня, таким образом, воздействие, скорее, оказывалось на наиболее нуждающихся, что нарушает SUTVA). Для того чтобы снизить влияние этой проблемы, можно произвести взвешивание по стратам, чтобы получить оценки, специфичные для страт.

Различающийся масштаб инвестиций предполагает, что при разном значении  $d$  (дозы воздействия) эффект будет различаться. Мы можем построить функцию дозу-отклик, если изменим наше воздействие с бинарного на непрерывное. Либо же мы можем получить оценки для разных уровней инвестиций, если запустим отдельные модели для маленького, среднего и высокого масштабов инвестиций. Еще одна альтернатива – запустить одну модель, где включить взаимодействие между воздействием ( $D$  – фактом наличия инвестиционного проекта) и масштабом инвестиций (разбить эту переменную на категории).

В идеале нам хотелось бы, чтобы количество муниципалитетов в двух группах было примерно равным. Особенность остро эта проблема может встать при мэтчинге, когда для и без того небольшой экспериментальной группы нужно будет отыскать пары в контрольной. В целом, чем больше муниципалитетов включено в исследование, тем выше будет статистическая мощность. Длительный период наблюдений также позволит учитывать временные эффекты, влияющие на доходы населения.

Я посчитал необходимым вначале ответить на последний вопрос в тестовом задании, потому что проблемы, поднятые выше, должны быть адресованы на этапе планирования исследования. Ниже же я предложу один из наиболее распространенных дизайнов для случая, когда воздействие было (квази-)случайным – разность-разностей (DiD). Мы можем замерить средний эффект воздействия на группу воздействия (ATT), используя вариацию между муниципалитетами, подвергшимися и не подвергшимися воздействию во времени. Я опишу базовый подход, но дополнительно сошлюсь на литературу по дизайну разность-разностей, когда воздействие наступает не одномоментно, а в разные временные периоды (см. литературу по Event-study DiD и Staggered DiD). Так, если предположение о параллельности трендов выполняется – в отсутствие воздействия тренды зависят от переменной в группах воздействия и контроля были бы одинаковыми, – то мы могли бы оценить эффект инвестиционных проектов на муниципалитеты, подвергшиеся воздействию, по сравнению с гипотетической ситуацией, когда эти же муниципалитеты не получили воздействие. Именно предположение о параллельности трендов позволяет нам использовать наблюдения в контрольной группе для расчета (или восстановления / импутации) контрфактуального тренда (контрфактуальных значений) в группе воздействия.