# Подготовка данных и описание методики

Согласно заданию, были выбраны наиболее актуальные данные Росстата о ключевых показателях регионов за 2017-2020 год. На их основе с использованием инструмента *Excel* была создана сводная таблица *Full\_data.xlsx* (См. в приложении).

В результате первичной обработки к данным были добавлены дополнительные признаки, отражающие изменение рассматриваемых показателей за выбранный период времени. Так же были добавлены признаки, отражающие отклонение показателей роста от среднего значения, рассчитанного по округу, к которому принадлежит рассматриваемая область. Добавленные признаки выделены в таблицу зеленым цветом.

Последующая обработка данных была выполнена в инструменте *Jupyter Notebook* с использованием библиотек *pandas* и *sklearn*. Методика сегментирования территорий и эффективности деятельности была основана на методе кластеризации *KMeans*, разделяющим объекты на основании подсчета их внутрикластерного расстояния в признаковом пространстве.

Перед проведением кластеризации была рассмотрена матрица корреляции Спирмена, отражающая зависимость между рассматриваемыми переменными (Рисунок см. в *ipynb*). По итогам анализа матрицы корреляции было выявлено, что положительно скоррелированы признаки, связанные с объемом строительства и количеством жителей, количеством жителей и относительным количеством высокопроизводительных мест труда, средней зарплатой и уровнем бедности, объемом строительства и количеством высокопроизводительных мест труда, приростом населения и уровнем бедности. Отрицательно скоррелированы признаки, связанные с количеством жителей и уровнем бедности, кол-вом высокопроизводительных мест труда и уровнем бедности. Так же большим коэффициентом корреляции обладают показатели роста показателей за смежные годы.

В связи с большим количеством скоррелированных признаков, перед проведением кластеризации было произведено снижение размерности рассматриваемого векторного пространства с использованием метода главных компонент (*PCA*). В каждом рассматриваемом случае размерность итогового пространства подбиралась для достижения 90% значения объясняемой дисперсии изначальных данных (Рисунок 1).

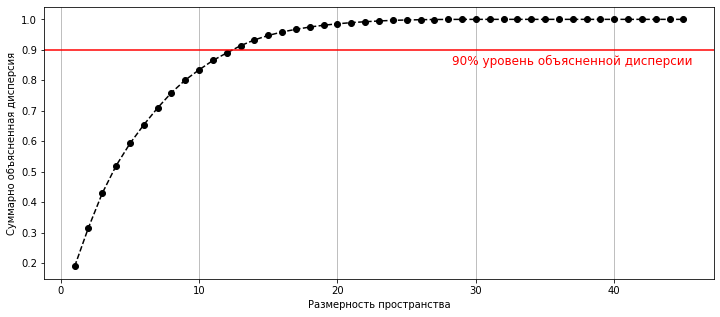
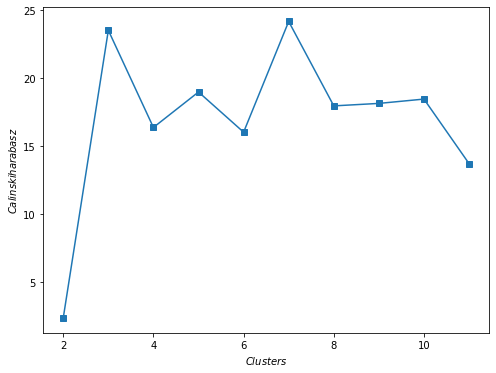


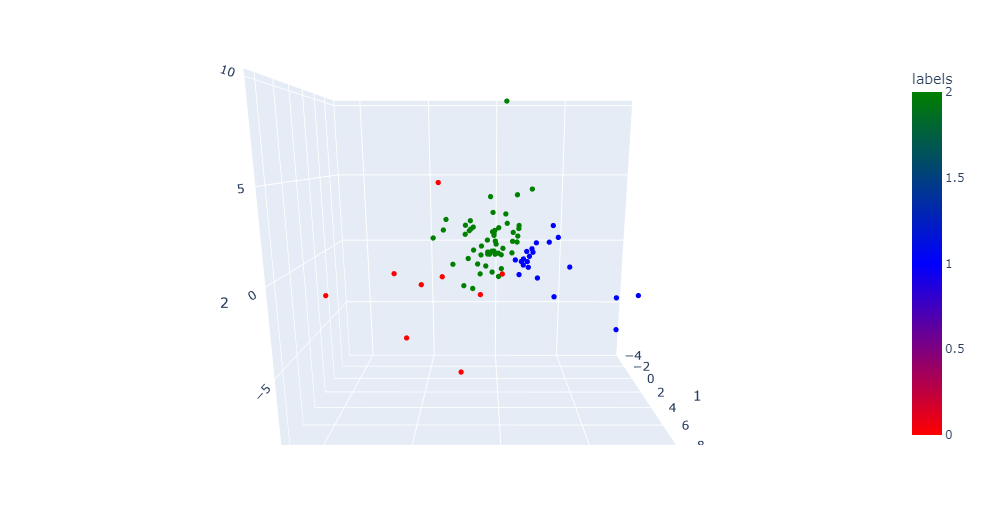
Рисунок 1 – Подбор размерности пространства для кластеризации

# Сегментирование территорий

Сегментирование территорий было проведено на основе кластеризации всех имеющихся данных, кроме численности населения. В качестве метрики для выбора количества кластеров для разбиения был выбран индекс Калински-Харабаса, показывающий отношение между межгрупповой дисперсией и внутригрупповой дисперсией. Проверка модели проходила итеративно.

 Рисунок 2 – Индекс Калински-Харабаса в зависимости от количества кластеров

На основании полученных данных регионы были разбиты на 3 кластера (0, 1, 2). Кластеры было решено назвать 1 - «Регионы с опережающим развитием», 2 – «Регионы со средним развитием», 0 - «Отстающие регионы». Отображение сегментированных регионов в пространстве главных компонент представлено на рисунке 3.



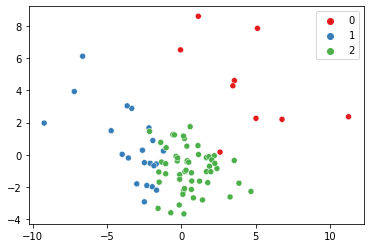


Рисунок 3 – Сегментированные территории в пространстве 2 и 3 главных компонент

К «Регионам с опережающим развитием» были отнесены Белгородская область, Воронежская область, Калужская область, Липецкая область, Московская область, Ярославская область, г. Москва, Ленинградская область, г. Санкт-Петербург, Краснодарский край, Волгоградская область, Ростовская область, Республика Башкортостан, Республика Татарстан,

Нижегородская область, Самарская область, Свердловская область, Тюменская область, Ханты-Мансийский автономный округ-Югра, Челябинская область, Новосибирская область.

К «Отстающим регионам» были отнесены Республика Калмыкия, Республика Дагестан, Республика Ингушетия, Кабардино-Балкарская Республика, Карачаево-Черкесская Республика, Чеченская Республика, Республика Алтай, Республика Тыва, Республика Саха (Якутия). На рисунке 3 отображены основные экономические показатели регионов.

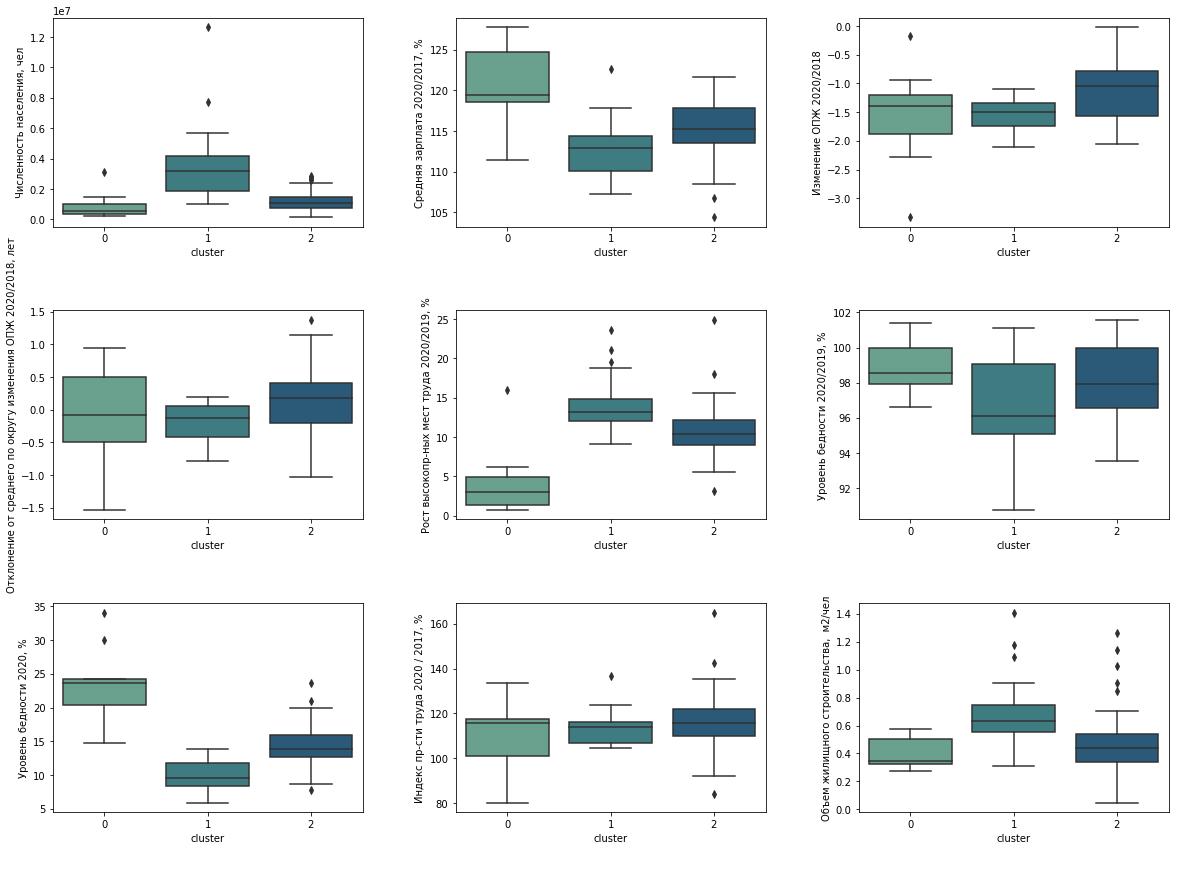


Рисунок 4 – Экономические показатели сегментированных регионов

В первую очередь «развитые» регионы характеризуются высокой численностью населения: 1.0 ÷ 12.6 млн человек, тогда как в «отстающих» проживает 0.22 ÷ 3.0 млн человек. Как видно из графиков, для «развитых» регионов характерно высокое количество высокопроизводительных мест труда, 0.12 ÷ 0.14 мест/чел против 0.02 ÷ 0.04 мест/чел, и значение роста высокопроизводительных мест труда за 2020/2019 год (12 ÷ 15% / 2 ÷ 4%).

Для «развитых» регионов характерно меньшее значение уровня бедности: 6 ÷ 12%, по сравнению с 21 ÷ 24 % для отстающих, а также более высокие темпы снижения бедности (95 ÷ 97%/98 ÷ 100%). Так же в «развитых регионах» более высокие показатели объемов жилищного строительства (0.6 ÷ 0.8 м2/чел против 0.3 ÷ 0.5 м2/чел). Для «отстающих» регионов характерно более высокое значение относительного роста средней зарплаты с 2017 по 2020 год: 111 – 128%, тогда как рост средних зарплат в «развитых» регионах составил 106 – 114%. Скорее всего, этот факт объясняется более низким базовым значением средней зарплаты в этих регионах.

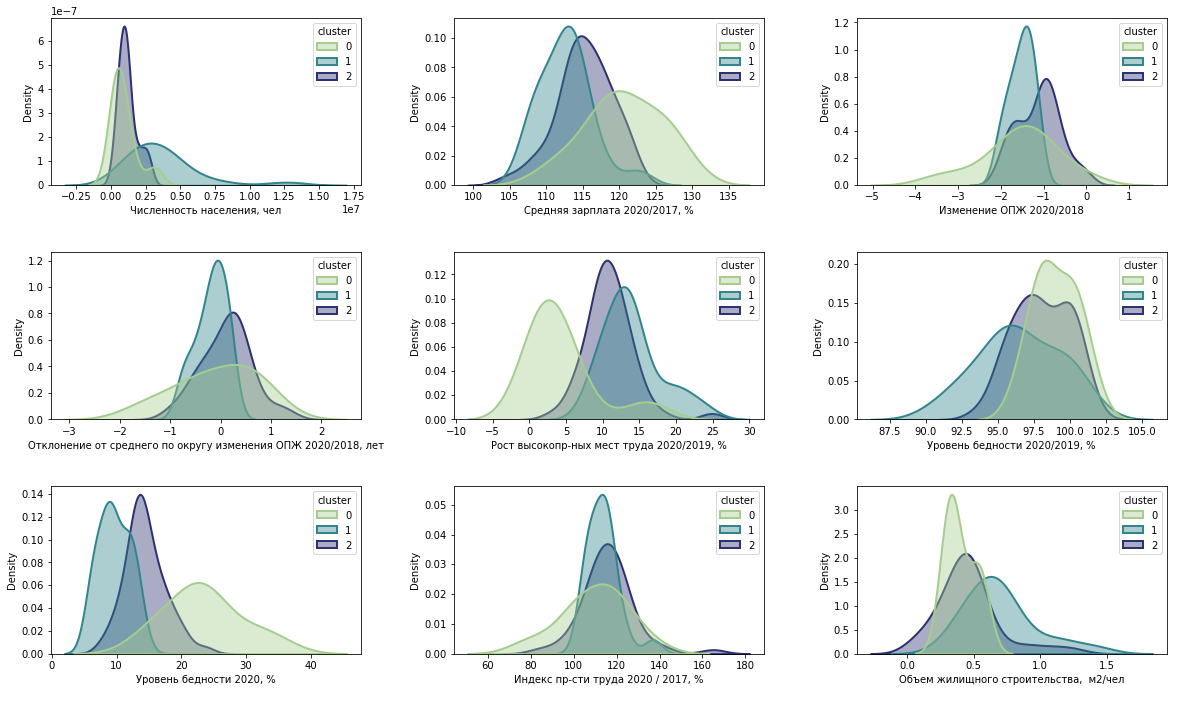


Рисунок 5 – Распределения экономических характеристик сегментированных территорий

Таким образом, кластер с «развитыми» территориями включает в себя регионы с высокой численностью населения, а так же большим количеством высокопроизводительных мест труда и низким уровнем бедности.

# Оценка эффективности деятельности губернаторов

Для оценки деятельности губернаторов была использована аналогичная модель кластеризации, но в изначальное признаковое пространство были включены только расчетные признаки, отражающие изменения целевых показателей в регионах по сравнению со средним значением изменения показателей в соответствующем округе. На основании максимального значения индекса Калински-Харабаса было выделено 5 кластеров, представленных на рисунке 6.

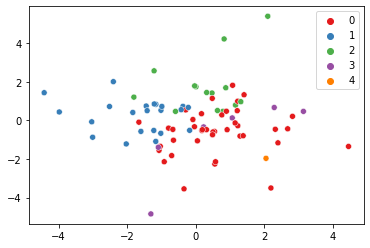


Рисунок 6 – Эффективность деятельности губернаторов в пространстве 2 главных компонент

Как видно из рисунка 6, один из кластеров состоит из одного региона (республика Коми) и будет рассмотрен отдельно.

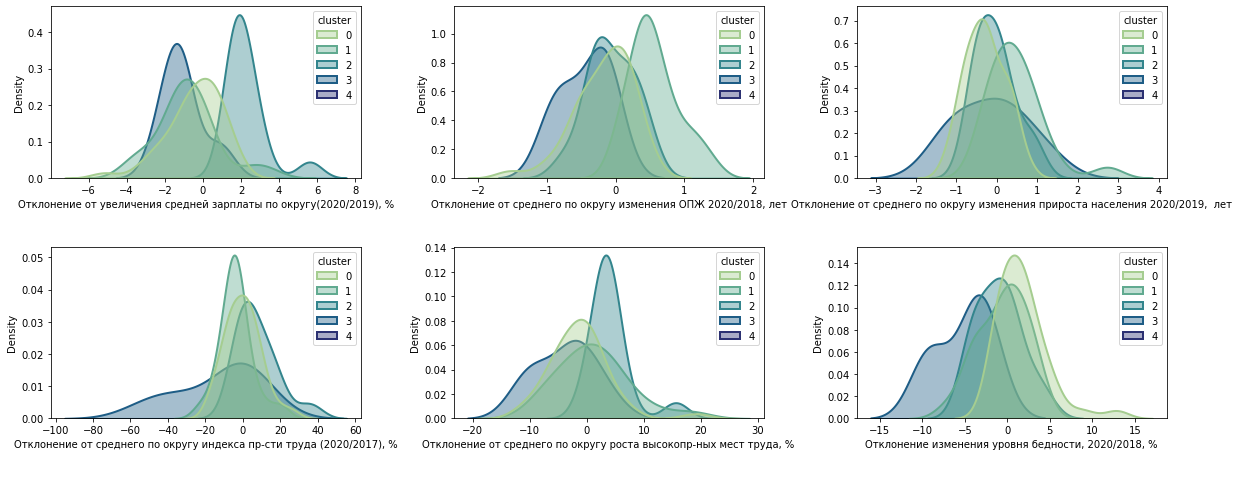


Рисунок 7 – Эффективность деятельности губернаторов из различных кластеров

На рисунке 7 наблюдается, что губернаторы, принадлежащие к кластеру 3, обладают наименьшими показателями роста показателей, определяющих эффективность деятельность высших должностных лиц, по сравнению с другими регионами из их округа. К этим регионам относятся: Калужская область, Московская область, г.Санкт-Петербург, Волгоградская область, Республика Саха (Якутия), Приморский край.

Для этого кластера характерны следующие значения :отклонение показателя увеличения заработной платы (2020/2019) находится в диапазоне -5.7 ÷ 0.6%, изменение ОПЖ (2020/2018) в диапазоне -0.95 ÷ 0.06 лет, отклонение по изменению прироста населения (2020/2019): -9.0 ÷ 0.2, отклонение по индексу производительности труда (2020/2017): -49.0 ÷ 7.4%, отклонение по уровню бедности (2020/2018): -9.4 ÷ -2.3%.

Республика Коми относится к 3 кластеру по всем показателям, кроме прироста населения и изменения уровня бедности и соответственно, эта территория так же может рассматриваться для возможной смены управленцев.

Таким образом построение автоматического алгоритма увольнения губернаторов возможно с использованием алгоритма кластеризации *KMeans,* возможно обучение на этих размеченных данных классификатора в виде логистической регрессии, дерева решений, градиентного бустинга. Так же возможно сравнивать новые данные с медианным значением выделенных показателей с предварительно определенным коэффициентом значимости для каждого фактора и разработать на этой основе специализированную систему скоринга.

*P.S.* На основании представленных данных была построена 3-я модель на основании признаков, отражающих рост представленных показателей. На основании этой модели регионы возможно разделить на более мелкие кластеры, чем представлены в пункте 1. Это позволяет выделить регионы, лидирующие в одной из представленных областей. Эта модель при необходимости может быть более подробно описана по дополнительному запросу.