МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ»
(национальный исследовательский университет)»

Институт (Филиал) № 8 «Компьютерные науки и прикладная математика» **Кафедра** 806

Группа М8О-408Б-20 **Направление подготовки** 01.03.02 «Прикладная математика и информатика»

Профиль Информатика

Квалификация: бакалавр

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

на тему: Кластеризация образовательных организаций с учетом контекстных данных методами машинного обучения

Автор ВКРБ:	Зубко Дмитрий Валерьевич	(_____)
Руководитель:	Левинская Мария Александровна	(_____)
Консультант:	—	(_____)
Консультант:	—	(_____)
Рецензент:	—	(_____)

К защите допустить

Заведующий кафедрой № 806 Крылов Сергей Сергеевич (_____)

_____мая 2024 года

Москва 2024

РЕФЕРАТ

Выпускная квалификационная работа бакалавра состоит из 48 страниц, 40 рисунков, 3 таблиц, 30 использованных источников и 1 приложения.

ОБРАЗОВАТЕЛЬНЫЕ ОРГАНИЗАЦИИ, МАШИННОЕ ОБУЧЕНИЕ, НОРМАЛИЗАЦИЯ, КЛАСТЕРИЗАЦИЯ, АНАЛИЗ, ТЕПЛОВЫЕ КАРТЫ, ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ.

Цель работы – разработать программное обеспечение для кластеризации образовательных организаций и последующего анализа полученных профилей (кластеров) образовательных организаций.

Основное содержание работы состояло в создании программного обеспечения для кластеризации образовательных организаций, что было связано с задачами подготовки данных и выбора метода кластеризации.

Основной результат работы – это профили образовательных организаций, их описание и программное обеспечение, позволяющие кластеризовать образовательных организации на основе диагностической работы НИКО и анкет учащихся со свободным ответом.

Результаты разработки предназначены для улучшения оценки качества в образовательной сфере, могут применяться в задачах рейтингования образовательных организаций и прогнозирования результатов образовательных организаций.

СОДЕРЖАНИЕ

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

В настоящей выпускной квалификационной работе бакалавра применяют следующие термины с соответствующими определениями:

Кластер – это группа объектов, объединенная по каким-то признакам в одну группу

Кластеризация – это задача разделения выборки объектов на кластеры, внутри которых объекты более похожи друг на друга, по сравнению с объектами из других кластеров

Датафрейм – это двумерная структура данных со столбцами и строками

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

В настоящей выпускной квалификационной работе бакалавра применяютследующие сокращения и обозначения:

ОО – образовательная организация

НИКО – национальное исследование качества образования

ФИОКО – Федеральный Институт Оценки Качества образован

ВПР – всероссийская проверочная работа

ВВЕДЕНИЕ

Данная работа является второй частью проекта «Кластеризация общеобразовательных организаций с учетом психологического портрета учащихся». Первая часть проекта включает разбор свободных ответов учащихся. В этой работе используются данные, полученные в первой части, такие как тип семьи и индексы, относящиеся к истории, культуре и семейной жизни учащихся.

В России создана Единая система оценки качества образования, известная как ЕСОКО. Эта система помогает отслеживать академические достижения учащихся на различных этапах школьного образования. Она также способствует своевременному обнаружению и решению возникающих проблем в образовательной сфере, учитывая предметные, школьные и региональные аспекты.

Эта система предоставляет полный обзор состояния образования в стране и способствует анализу разнообразных аспектов, оказывающих влияние на результаты деятельности школ. Она дает возможность школьным учреждениям самостоятельно оценивать свою деятельность и определять существующие проблемы. Кроме того, она предоставляет родителям данные о качестве знаний их детей.

Система оценки качества школьного образования в России состоит из: единого государственного экзамена (ЕГЭ), основного государственного экзамена (ОГЭ). Промежуточные срезы знаний проводятся при помощи разных работ, например, НИКО и ВПР.

Центр проведения национальных и международных исследований качества образования, который входит в состав ФИОКО, реализует на территории России НИКО, а также организует проведение всероссийских проверочных работ.

Актуальность данной работы связана с тем, что ФИОКО сможет её использовать для автоматизации анализа НИКО и ВПР.

Цель работы – применении методов машинного обучения для кластеризации образовательных организаций на основе контекстный данных с последующим анализом полученных профилей образовательных организаций. Для достижения поставленной цели в работе были решены следующие задачи:

– преобразование данных об учащихся в образовательных организациях в данные об образовательных организациях;

– дополнение данных об образовательных организациях;

– нормализация и очистка данных;

– применение выбранного метода кластеризации к данным и создание «профилей» (кластеров) образовательных организаций;

– поиск корреляций полученных профилей образовательных организаций с успеваемостью учащихся;

– создание тепловых карт для каждого кластера.

Программное обеспечение состоит из 5 файлов:

– diploma_preparing_dataset_people для обработки и нормализации данных об учащихся;

– diploma_preparing_dataset_schools для создания датасета с характеристиками образовательных организаций;

– diploma_clustering для кластеризации образовательных организаций;

– diploma_analysis для анализа данных;

– diploma_russian_map для построения тепловых карт.

1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Профили образовательных организаций

Профили образовательных организаций представляют собой описание образовательных организаций, выделенных в различные кластеры на основе контекстных данных.

Имея профили образовательных организаций, предоставляется множество возможностей для работы и поддержки этих образовательных организаций:

– анализ и улучшение учебных планов: используя данные об успеваемости учащихся образовательных организаций, можно определить, какие предметы или темы требуют дополнительного внимания или изменения в методах преподавания;

– профессиональное развитие учителей: анализируя, какие образовательные организации имеют наилучшие результат, можно организовывать обмен опытом и наставничество среди педагогического состава;

– прогнозирование успеваемости: используя аналитические инструменты, можно прогнозировать будущую успеваемость учащихся в выбранной образовательной организации, что позволяет заранее предотвращать возможные учебные трудности;

– распределение ресурсов: данные о профилях образовательных организаций могут помочь определить, в каких областях и классах может потребоваться дополнительная поддержка, что позволяет более эффективно распределять ресурсы и помощь;

– выявление резильентных школ: резильентные или устойчивые школы – это образовательные учреждения, которые показывают хорошие учебные результаты несмотря на то, что они расположены в экономически неблагополучных районах или сталкиваются с другими внешними трудностями. Сбор и анализ данных о профилях школ могут помочь выявить такие учреждения и понять, какие методы и подходы к управлению и преподаванию они используют. Это позволит распространить успешные практики среди других школ, а также дать возможность для проведения целенаправленных исследований эффективности этих методов.

1.2 НИКО

Существует актуальная задача создания сбалансированной системы процедур оценки качества общего образования. Эта система должна обеспечивать надежную информацию о состоянии различных компонентов региональных и муниципальных образовательных систем, их соответствии требованиям ФГОС и общей оценке состояния системы образования в России. Для достижения этой цели наиболее эффективным способом является проведение регулярных исследований качества образования, основанных на сборе и анализе разнообразных данных о состоянии региональных и муниципальных систем.

С 2014 года в России реализуется программа НИКО [19]. В рамках НИКО проводится регулярное исследование качества преподавания отдельных учебных предметов на различных уровнях образования, проводимое не реже двух раз в год. Каждое исследование рассматривается как самостоятельный проект. Целями программы НИКО являются:

– укрепление единого образовательного пространства в России;

– поддержка выполнения поручений Президента РФ и программных документов Правительства РФ, касающихся качества образования;

– улучшение механизмов получения точной и содержательной информации о состоянии различных уровней и подсистем образовательной системы, включая внедрение ФГОС;

– развитие информационно-аналитической и методологической базы для принятия управленческих решений по развитию образования в России;

– содействие эффективному внедрению ФГОС;

– поддержка процессов стандартизации процедур оценки в сфере образования.

Согласно стратегии научно-технологического развития России, основная цель заключается в обеспечении независимости и конкурентоспособности страны путем создания эффективной системы, которая максимально раскрывает и использует интеллектуальный потенциал нации. Для этого необходимо создать условия для выявления талантливой молодежи и поддержки их карьерного роста в области технологий и инноваций. В этом контексте важным шагом является проведение Национальных исследований

качества образования, направленных на оценку личностных и метапредметных достижений учеников 6 и 8 классов.

Целью исследования НИКО является оценка соответствия результатов обучения в основной школе установленным ФГОС, включая личностные и метапредметные аспекты; выявление текущих проблем, требующих решения на федеральном, региональном, муниципальном и школьном уровнях через программы воспитательной работы; а также разработка рекомендаций для всех участников образовательного процесса [20].

Задачами исследования НИКО являются:

– всесторонняя диагностика уровня подготовки учащихся 6 и 8 классов общеобразовательных школ;

– сбор, обработка и анализ данных об образовательном процессе в ОО;

– подготовка аналитических отчетов на основе полученных данных;

– разработка рекомендаций по применению результатов исследования;

– приведение обсуждений результатов исследования.

1.3 База данных НИКО

Ежегодно проводятся исследования НИКО для учащихся 6-8 классов. В данной работе используется база НИКО 2021-2022, в которой не содержатся персональные данные учащихся. На рисунке 1 показана часть имеющейся базы.

№	Регион	Класс	Пол	РУ Теку	МА Тек	Расскажите немного о своей семье. Кто в неё входит?	Кем из членов семьи Вы гордитесь? Объясните, почему.	Какие традиции существуют в Вашей семье?
1964895	78	6	Девочки	3	3	бабушка мама папа бра	мама потому что об	помогать друг дру
1965099	78	6	Девочки	4	4	мама,папа,сестра	мамой и сестрой.мама-	стала писател
1964302	78	6	Девочки	5	5	я, папа, мама, брат, бра	мамой, потому что	на день рождения
1965852	78	6	Мальчики	4	4	Мама,Папа,собака,я	Я горжусь всеми чле	покупать что нибу
1953331	78	6	Мальчики	3	3	Папа,мама, две сестры,	Папа он очень мног	В моей семье трад

Рисунок 1 – Часть базы НИКО

Свободные ответы учащихся были разобраны в первой части проекта.

1.4 Python и его библиотеки

1.4.1 Python

Python – это высокоуровневый, интерпретируемый язык программирования с динамической типизацией, автоматическим управлением памятью и мощной стандартной библиотекой [1]. Он был создан Гвидо ван Россумом и впервые выпущен в 1991 году.

Язык ориентирован на повышение продуктивности разработчика и читаемости кода. Python поддерживает несколько парадигм программирования, включая структурное, объектно-ориентированное и функциональное программирование.

Python часто используется для веб-разработки, науки о данных, создания скриптов, автоматизации и множества других областей. Он известен своим ясным синтаксисом, который часто описывается как похожий на читаемый человеком английский текст, что делает Python популярным выбором

1.4.2 Pandas

Pandas – это открытая библиотека для языка программирования Python, предназначенная для обработки и анализа данных [2]. Она предоставляет специализированные структуры данных и операции для манипуляции с числовыми таблицами и временными рядами. Библиотека pandas является одним из самых популярных и полезных инструментов в сфере data science и анализа данных на Python, благодаря своей эффективности и простоте в использовании.

1.4.3 Matplotlib

Matplotlib – это библиотека для создания статических, интерактивных и анимированных визуализаций в Python [3]. Она была создана Джоном Хантером в 2003 году и с тех пор стала стандартным инструментом для построения графиков и диаграмм в научных и инженерных исследованиях.

Библиотека предлагает широкий набор инструментов для построения различных типов графиков, включая:

– линейные графики;

– диаграммы рассеяния;

– столбчатые диаграммы;

– гистограммы;

– тепловые карты.

Matplotlib проектировался с прицелом на создание публикаций качественных изображений в различных форматах, но при этом он также поддерживает интерактивное отображение графиков в различных средах, таких как Jupyter notebooks.

Библиотека highly configurable, что позволяет пользователям настраивать практически все аспекты фигур и осей, включая тип и цвет линий, шрифты, легенды, насечки на осях и многое другое. Часто для удобства работы с Matplotlib используются дополнительные библиотеки, такие как Pandas и Seaborn, которые обеспечивают более высокоуровневые абстракции для создания сложных визуализаций.

1.4.4 Seaborn

Seaborn – это библиотека для создания статистических графиков в Python [4]. Она построена на основе Matplotlib, однако предоставляет более высокоуровневый интерфейс для рисования красивых и информативных статистических графиков. Seaborn облегчает создание типовых графиков: диаграмм рассеяния, гистограмм, коробчатых диаграмм, схем вулканов и многих других. Кроме того, Seaborn хорошо интегрируется с Pandas, что упрощает работу со сложными структурами данных и проведение статистического анализа данных. Эта библиотека также автоматически настраивает графический стиль и цветовые схемы графиков, чтобы они были не только информативными, но и приятными для глаза.

1.4.5 Sklearn

Scikit-learn, часто сокращенно называемый sklearn, – это библиотека для машинного обучения, написанная на языке программирования Python [5]. Она включает в себя широкий спектр алгоритмов и инструментов для моделирования и анализа данных, таких как классификация, регрессия,

кластеризация и выбор признаков.

Scikit-learn хорошо известна своим простым и единообразным программным интерфейсом, ориентированным на повседневные задачи анализа данных. Она использует другие библиотеки, такие как NumPy и SciPy, для численных вычислений и математических операций, и предполагает, что пользователи имеют базовое понимание машинного обучения.

1.4.6 GeoPandas

GeoPandas – это популярная библиотека на языке Python, предоставляющая инструменты для работы с геопространственными данными [6]. Она расширяет возможности библиотеки pandas, добавляя тип данных «гео» для пространственных операций на геометрических типах данных. Это делает GeoPandas удобным инструментом для анализа и визуализации геопространственной информации.

Умеет использовать GeoJSON формат [7]. GeoJSON — это формат обмена геопространственными данными, который использует JSON (JavaScript Object Notation) для кодирования различных геометрических структур и их атрибутов. В GeoJSON можно представлять такие элементы, как точки, линии, полигоны, а также более сложные структуры на основе этих примитивов. Этот формат хорошо адаптирован для использования в интернете, так как он основан на JSON, который широко поддерживается современными веб-технологиями и является легко читаемым как для человеков, так и для машин. Основные элементы формата:

– объект (Feature) – основная единица данных, представляет собой географический объект с геометрией и дополнительными свойствами;

– геометрия (Geometry) – описывает форму объекта, может быть нескольких типов: точка (Point), линия (Linestring), полигон (Polygon), множественные точки (MultiPoint), множественные линии (MultiLineString), множественные полигоны (MultiPolygon), коллекция геометрий (GeometryCollection);

– коллекция объектов (FeatureCollection) – содержит массив объектов (features).

На рисунке 2 показан пример структуры точки.

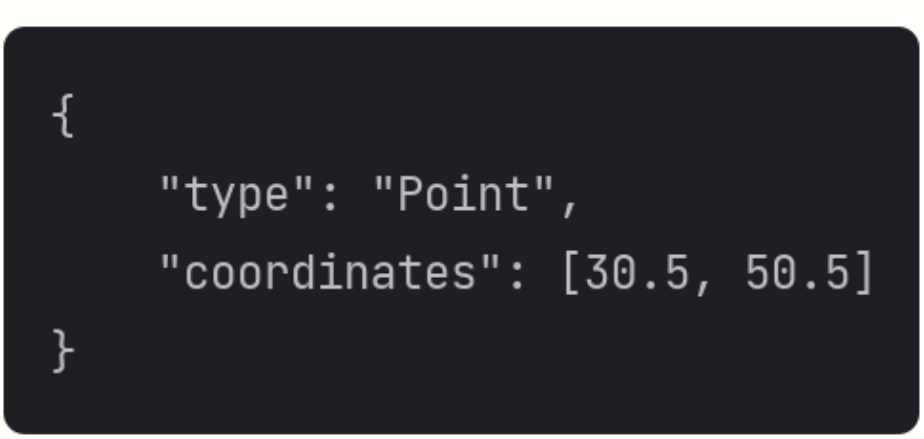

Рисунок 2 – Объект точка

На рисунке 3 показан пример, как описаны регионы России с помощью формата geojson.

Рисунок 3 – Описание регионов России

1.4.7 Folium

Folium – это библиотека для Python, которая используется для визуализации данных на картах [8]. Она основана на библиотеке Leaflet.js, которая позволяет создавать интерактивные карты в браузере. Folium упрощает процесс создания карт на основе данных, доступных в Python, предоставляя высокоуровневый интерфейс для Leaflet, который наследует его мощные возможности для визуализации геопространственной информации.

1.5 Машинное обучение

Машинное обучение (Machine Learning, ML) – это подраздел искусственного интеллекта (AI), который фокусируется на разработке

алгоритмов и статистических моделей, позволяющих компьютерам улучшать свои задачи посредством опыта и данных, а не только за счёт явного программирования. То есть это подход, при котором системы способны обучаться и улучшать свои действия, анализируя большие объемы данных. В машинном обучении используются разнообразные типы алгоритмов, каждый из которых имеет свои особенности и подходит для решения определенного круга задач.

1.5.1 Z-нормализация

Z-нормализация, также известная как стандартизация, является статистическим методом нормализации атрибутов данных, целью которого является придание исходному распределению данных среднего значения равного 0 и стандартного отклонения равного 1. Это достигается путём вычитания среднего значения каждого признака из каждого значения признака и последующим делением на стандартное отклонение всего признака.

Формула для z-нормализации [9] атрибута x выглядит следующим образом:

$$z = \frac{(x - \mu)}{\sigma}, \quad (1)$$

где:

– x – значение атрибута;

– μ (мю) – среднее значение атрибута в наборе данных линейные графики;

– σ (сигма) – стандартное отклонение атрибута в наборе данных;

– z – результат z-нормализации данного значения атрибута.

После преобразования каждое значение в наборе данных будет иметь своё собственное z-значение, которое отражает количество стандартных отклонений, на которое это значение отстоит от среднего. Такие преобразованные данные облегчают сравнение значений с разными масштабами и полезны во многих методах машинного обучения, особенно тех, что чувствительны к масштабу атрибутов данных.

1.5.2 K-Means

Существуют различные методы кластеризации [27]. K-means – это популярный алгоритм кластеризации, который используется в области машинного обучения и анализа данных для разделения данных на группы (кластеры) на основе подобия их элементов [10]. Основная идея алгоритма k-means заключается в разделении набора данных на k предварительно заданных групп (или кластеров), минимизируя сумму квадратов расстояний от точек до центров этих групп. Центр кластера (центроид) отражает среднее положение всех точек в кластере. По сути, алгоритм пытается оптимизировать расположение точек и центроидов так, чтобы внутрикластерное расстояние было как можно меньше, а расстояние между различными кластерами – как можно больше. Это осуществляется через итеративный процесс присваивания точек к ближайшим центроидам и обновления положения центроидов на основе текущего состава кластера.

1.5.3 Иерархическая кластеризация

Иерархическая кластеризация – это метод анализа данных, который строит иерархию кластеров, или групп, схожих элементов на основе их характеристик или расстояний между ними [11]. В контексте иерархической кластеризации, метод «дальнего соседа» (также известный как метод комплетной связи) – это одна из стратегий определения расстояния между кластерами.

В методе дальнего соседа расстояние между двумя кластерами определяется как максимальное расстояние между любой парой элементов, где один элемент принадлежит одному кластеру, а другой элемент другому кластеру. Проще говоря, для каждой пары кластеров, вы измеряете расстояния между всеми парами элементов (один из первого кластера, другой из второго кластера) и выбираете наибольшее из них как расстояние между кластерами.

Алгоритм иерархической кластеризации с использованием метода дальнего соседа следующий:

– сначала рассматривается каждый объект как отдельный кластер;

– рассчитываются расстояния между всеми парами кластеров в соответствии с выбранным методом (в данном случае методом дальнего соседа);

– соединяются два кластера, расстояние между которыми наименьшее, чтобы сформировать новый кластер;

– расстояния между новым созданным кластером и остальными кластерами обновляются таким образом, что для пары кластеров теперь используется максимальное расстояние между элементами, как было описано ранее;

– последние два шага повторяются до тех пор, пока все объекты не будут объединены в один единственный кластер.

1.6 Диаграммы

1.6.1 Ящичковая диаграмма

Ящичковая диаграмма – это статистический график, который используется для визуализации распределения числовых данных и их вариативности [21]. Она показывает пять ключевых числовых характеристик данных: минимальное значение, первый квартиль (Q1), медиану (Q2), третий квартиль (Q3) и максимальное значение. Ящичковая диаграмма позволяет быстро оценить центральную тенденцию, разброс и наличие выбросов в наборе данных. Содержит следующие компоненты: коробка, медиана, усы, выбросы.

Коробка простирается от первого квартиля (Q1) до третьего квартиля (Q3). Она охватывает интерквартильный размах (IQR), который содержит средние 50% данных.

Линия внутри коробки указывает медиану (Q2) набора данных. Медиана разделяет данные на две равные половины.

Линии, идущие от коробки, представляют диапазон данных. Усы обычно простираются до самого маленького и самого большого значения, не являющихся выбросами, или до 1.5 * IQR от Q1 и Q3.

Точки, которые находятся за пределами усов, считаются выбросами и обозначаются отдельными точками или символами.

На рисунке 4 показан пример ящичковой диаграммы с перечисленными компонентами.

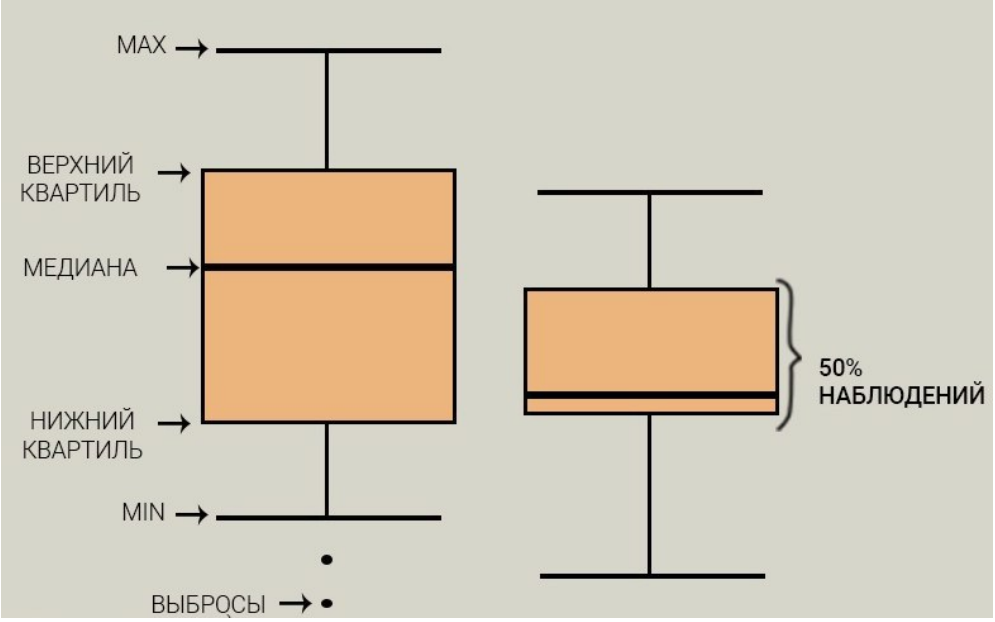

Рисунок 4 – Ящичковая диаграмма

1.6.2 Диаграмма рассеяния

Диаграмма рассеяния – это графическое представление двух непрерывных переменных, которое используется для определения связи между ними [22]. В этой диаграмме каждая точка представляет собой отдельное наблюдение и отображает значения обеих переменных. Главная цель диаграммы рассеяния показать, есть ли какая-либо взаимосвязь между двумя переменными, и если да, то какая она: положительная, отрицательная или отсутствует. На рисунке 5 показан пример диаграммы.

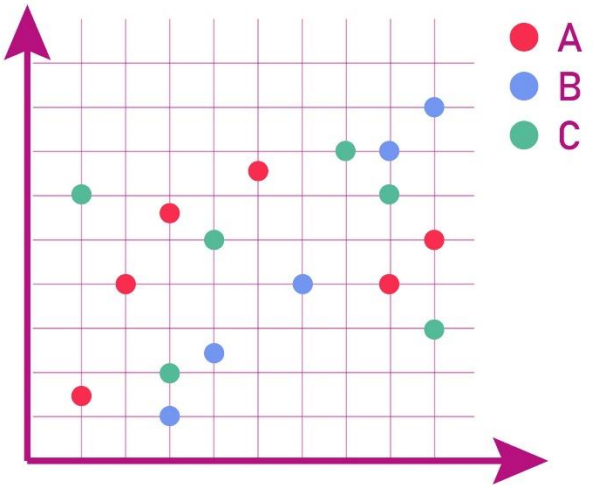

Рисунок 5 – Диаграмма рассеяния

1.7 Техническое задание

Задачей является разработать программное обеспечение для кластеризации образовательных организаций, используя данные диагностической работы учащихся. Применить разработанное программное обеспечение к диагностической работе, получить профили образовательных организаций. Проанализировать и описать полученные кластера.

2 РАЗРАБОТКА ПРОГРАММНОГО ПРОДУКТА

2.1 Предобработка имеющихся данных об учащихся в образовательных организациях

Данные изначально распределены по нескольким датасетам, данные в датасетах неперсонифицированы. Нужно совместить эти датасеты в единый датасет, учитывая только те поля, которые необходимы для проведения кластеризации и последующего исследования.

После объединения датасетов с помощью python и pandas проведена очистка данных. Чтобы не проводить дополнительные проверки, удалено немногочисленное количество записей, в которых тип семьи содержит значение «Нет четкого ответа» или нет отметки по математике или русскому языку.

Для просмотра средней зарплаты учителей по кластерам, добавлена средняя зарплата учителей. Для этого были взяты данные об средних зарплатах из открытого источника [12] и добавлены в датасет.

После просмотра датасета было обнаружено, что данных об учащихся в Северно-Кавказском федеральном округе и Южном федеральном округе мало. Для решения этой проблемы данные из Северно-Кавказского федерального округа были добавлены в Южный федеральный округ. Добавлены в Южный федеральный округ, потому что 19 января 2010 года Северно-Кавказский федеральный округ был отделен от Южного федерального округа [13].

Полученный датасет сохраняется в отдельный файл data_peaple.csv. В нем содержится информация о полях: Number, Логин_ОО, Class, Sex, Rus, Math, ПБ, Family, District, Region, Salary, History Knowledge Index, Famile Life index, Culture Index.

2.2 Преобразование данных об учащихся в данные об образовательных организациях

После получения всех необходимых данных учащихся необходимо описать образовательные организации. Перед этим удаляются данные об учащихся, в школах которых меньше 10 учащихся, чтобы полученные

значения были более точными. На рисунке 6 показан код, с помощью которого проводилась фильтрация.

```
df = df.groupby('Логин_ОО').filter(lambda x: len(x) >= 10)
```

Рисунок 6 – Удаление записей об учащихся, о школах которых мало информации

Для просмотра количества данных об образовательных организациях построена тепловая карта, в которой явно видно, сколько есть данных в округах и регионах. Всего в выборке присутствует 443 школы. Для построения тепловой карты данные группируются по двум полям: округу и региону. Сама тепловая карта строится при помощи библиотеки seaborn. На рисунке 7 изображена полученная тепловая карта.

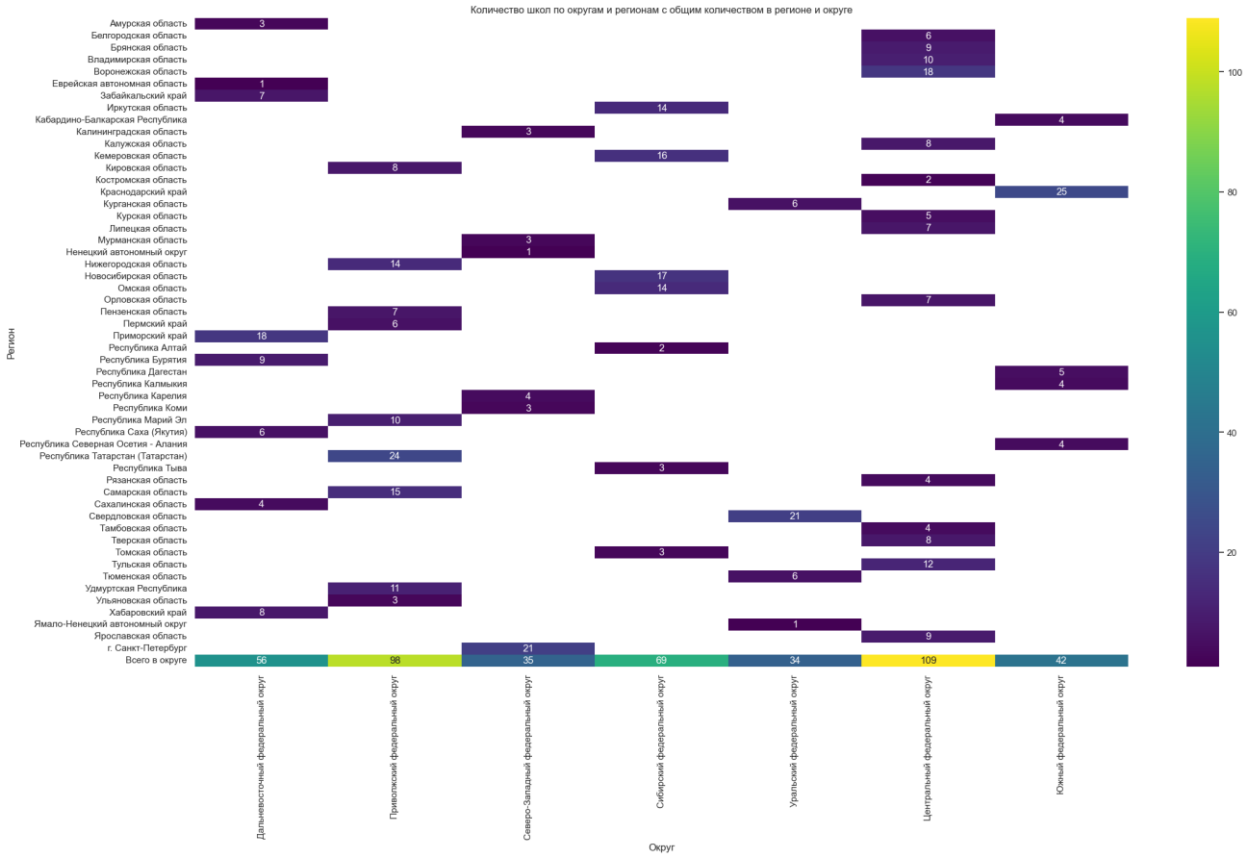

Рисунок 7 – Тепловая карта количества учащихся в округах и регионах

После просмотра тепловой карты можно сделать вывод, что имеются данные о многих регионах Росссии.

Для создания датасета, состоящего из образовательных организаций, нужно охарактеризовать школы по разным показателям, а затем объединить эти показатели в один датасет.

Первыми посчитаны средние оценки по математике и русскому языку и

первичный балл за диагностическую работу. Для этого данные группируются по логину образовательной организации, как показано на рисунке 8.

```
average_marks_schools_df = average_marks_schools_df.groupby('Логин_ОО').mean()
```

Рисунок 8 – Группировка по логину ОО и получение средних значений по школе по русскому языку, математике и первичному баллу

После группировки данных по логину образовательной организации и применения агрегирующей функции mean получился новый датасет, представленный таблицей 1.

Таблица 1 – Средние значения оценок по математике и русскому и первичного балла

Логин_ОО	Math	Rus	ПБ
edu030331	3.727273	3.818182	8.500000
edu030343	3.578947	3.736842	11.894737
…	…	…	…

Вторыми посчитаны средние значения индексов истории, культуры и семейной жизни, результат показан в таблице 2.

Таблица 2 – Средние значения индексов по истории, семейной жизни, культуры

Логин_ОО	Average History Knowledge Index	Average Family Life Index	Average Culture Index
edu030331	1.818182	1.636364	1.909091
edu030343	1.947368	1.842105	2.789474
…	…	…	…

Третьими посчитаны отношения типа семьи. Для этого применены агрегирующие функции во время группировки данных. Необходимо знать количество всех учащихся, учащихся с неполной, полной, многодетной полной семьями. На рисунке 9 показан код группировки при помощи метода groupby [29], с применением агрегирующих функций [23].

```
grouped_family_df = family_schools_df.groupby('Логин_ОО').agg(
    {
        'Неполная': 'sum',
        'Полная': 'sum',
        'Многодетная полная': 'sum',
        'Family': 'size'
    }
)
```

Рисунок 9 – Группировка по логину ОО и применение агрегирующих функций для подсчета количества разных типов в ОО

После группировки и применения агрегирующих функций, подсчитываются отношения. Для этого необходимо поделить количество семей каждого типа на общее количество учащихся. При этом сумма всех типов семей для каждой школы должна равняться единице. Результат отображен в таблице 3.

Таблица 3 – Отношения разных типов семей

Логин_ОО	Отношение_неполных	Отношение_полных	Отношение_многодетных_полных
edu030331	0.090909	0.681818	0.227273
edu030343	0.368421	0.631579	0.000000
…	…	…	…

Далее промежуточные значения объединяются в один датасет и сохраняются в отдельный файл data_school.csv. Объединение происходит при помощи функции merge [14] из библиотеки pandas. Также в итоговый датасет добавляются область, регион и средняя заработная плата учителей в регионе. Результирующий датасет содержит следующие поля: Логин_ОО, District, Region, Salary, Average History Knowledge Index, Average Culture Index, Average Culture Index, Отношение_неполных, Отношение_полных, Отношение_многодетных_полных, Rus, Math, ПБ.

2.3 Кластеризация образовательных организаций

Перед кластеризацией образовательных организаций построены графики, показывающие медианное значение среднего значения индексов истории, культуры и семейной жизни для каждого округа. Медианное значение – это значение, при котором половина данных находится выше него, а половина ниже [24]. Полученные графики отображены на рисунках 10-12.

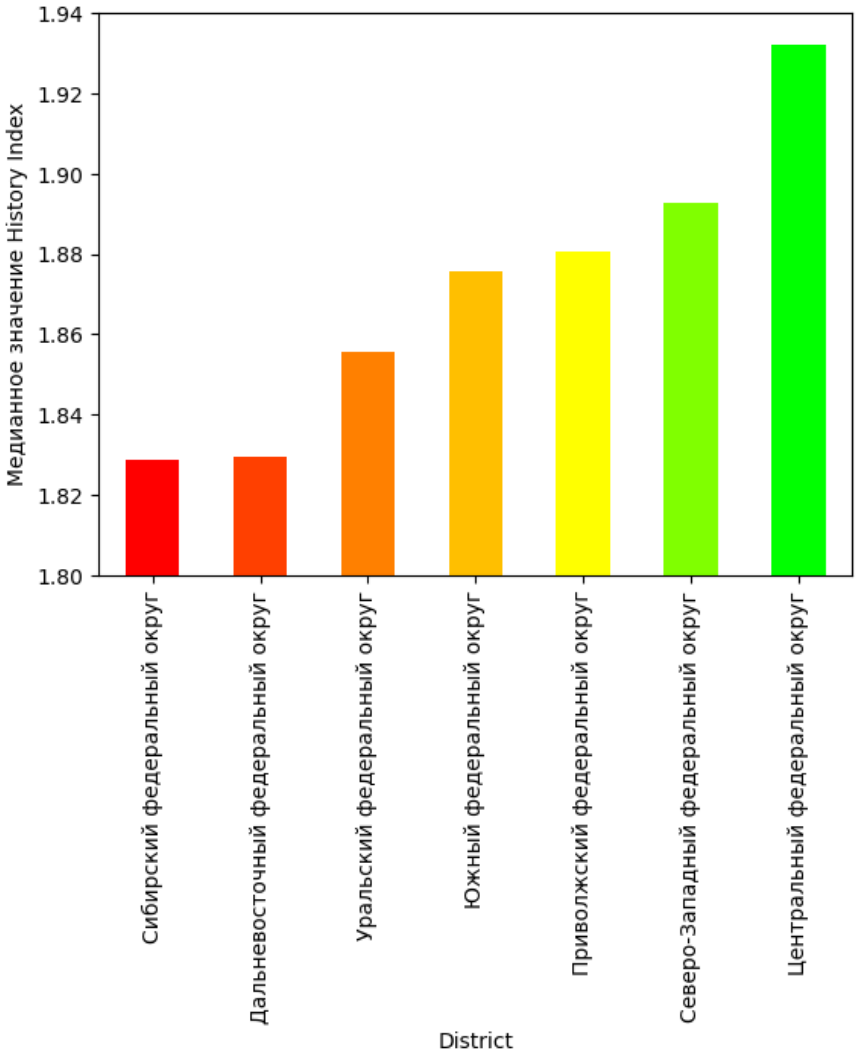

Рисунок 10 – Медианные значения индекса истории по округам

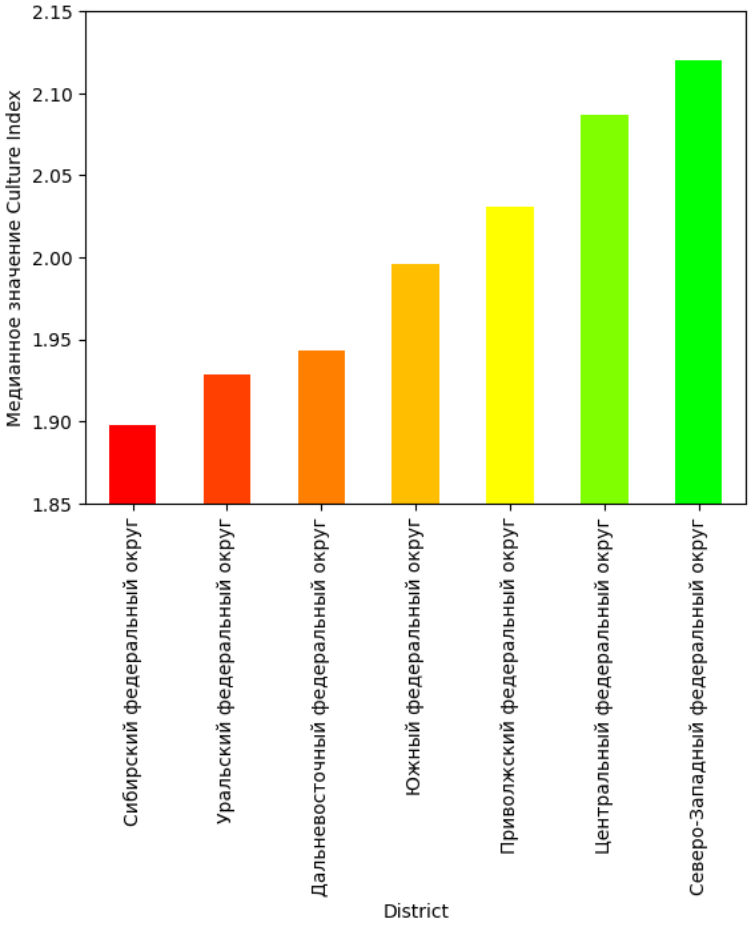

Рисунок 11 – Медианные значения индекса культуры по округам

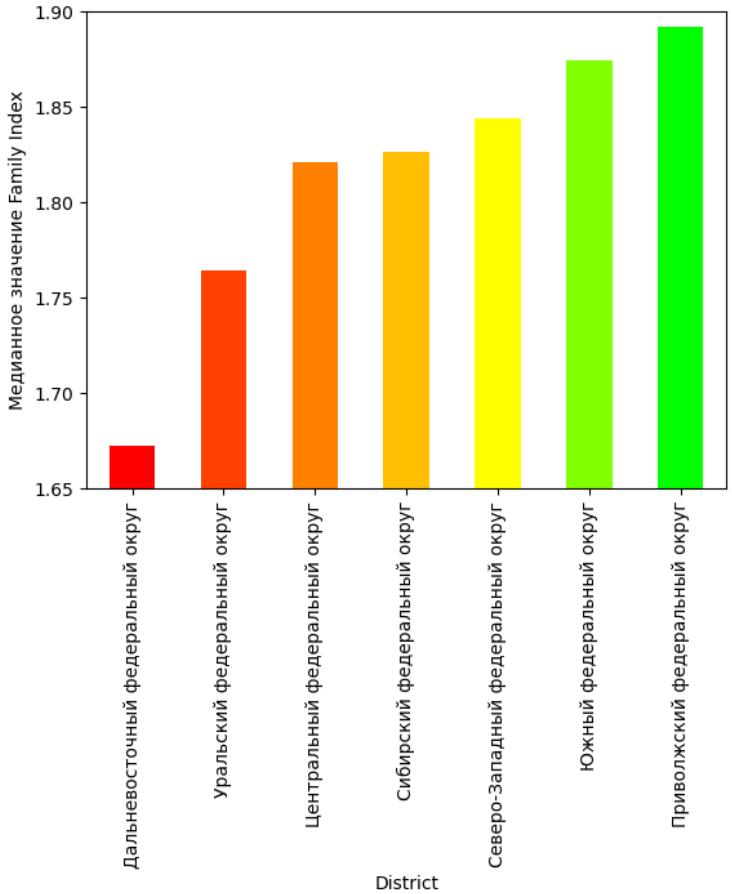

Рисунок 12 – Медианные значения индекса семейной жизни по округам

После просмотра полученных графиков видно, что Центральный федеральный округ лидирует по значению индекса истории, а Северно-Западный федеральный округ по значению индекса культуры. До просмотра графиков так и предполагалось, потому что в Центральном федеральном округе находится большое количество исторический объектов и располагается столица России – Москва. В Северно-Западном федеральном округе находится Санкт-Петербург, который считается культурной столицей России [15].

Для кластеризации образовательных организаций из датасета выделены поля, по которым происходит кластеризация: Average History Knowledge Index, Average Culture Index, Average Family Life Index, Отношение_неполных, Отношение_полных. Отношение_многодетных_полных. Остальные поля, такие как оценки по математику и русскому языку и первичный балл за диагностическую работу не участвуют в кластеризации.

Далее проведена нормализация данных Average History Knowledge Index, Average Culture Index, Average Family Life Index. Для нормализации этих данных используется StandartSkaler [9] из библиотеки sklearn. Этот скейлер реализует алгоритм z-нормализации.

Нормализация данных перед кластеризацией становится необходимой в случаях, когда признаки имеют разный масштаб или разные единицы измерения. Нормализация помогает уравнять важность всех признаков, что особенно важно для алгоритмов кластеризации, таких как K-means или иерархической, где расстояния между точками данных используются для определения принадлежности к кластерам.

После нормализации была построена визуализация алгоритма TSNE [16] из библиотеки sklearn. TSNE (t-distributed Stochastic Neighbor Embedding) – это алгоритм машинного обучения для визуализации, разработанный Лауренсом ван дер Маатеном и Джеффри Хинтоном. Он широко используется для уменьшения размерности больших многомерных наборов данных в 2D. Основная задача TSNE – отобразить объекты с высокоразмерного пространства на плоскость так, чтобы похожие объекты были отображены близко друг к другу, а непохожие – далеко друг от друга.

Из рисунка 8 видно, что среди данных можно выделить группы (кластера) похожих объектов.

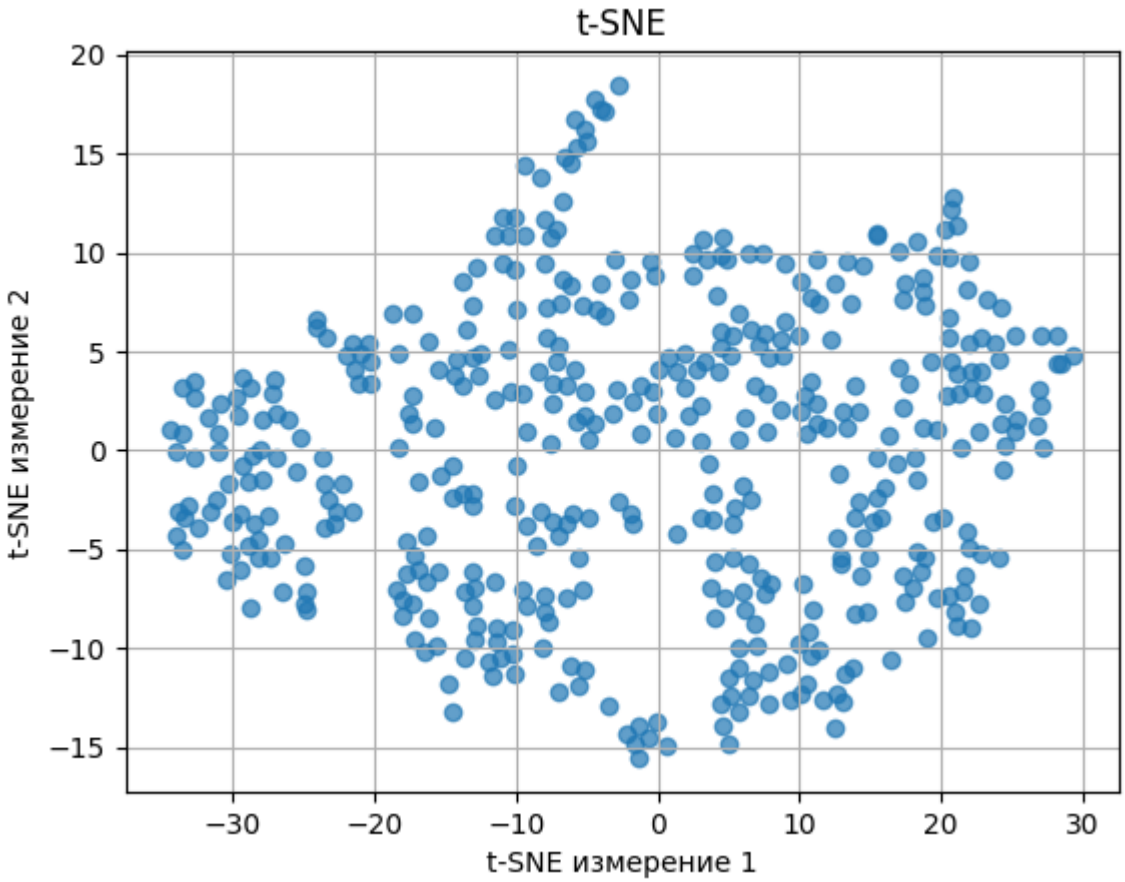

Рисунок 13 – Проекция данных в двумерное пространство

В первой попытке кластеризовать данные использовался алгоритм K-means [10]. Для определения оптимального количества кластеров в этом алгоритме использовался метод локтя [17]. На рисунке 14 показан график с средней квадратичной ошибкой для разного количества кластеров.

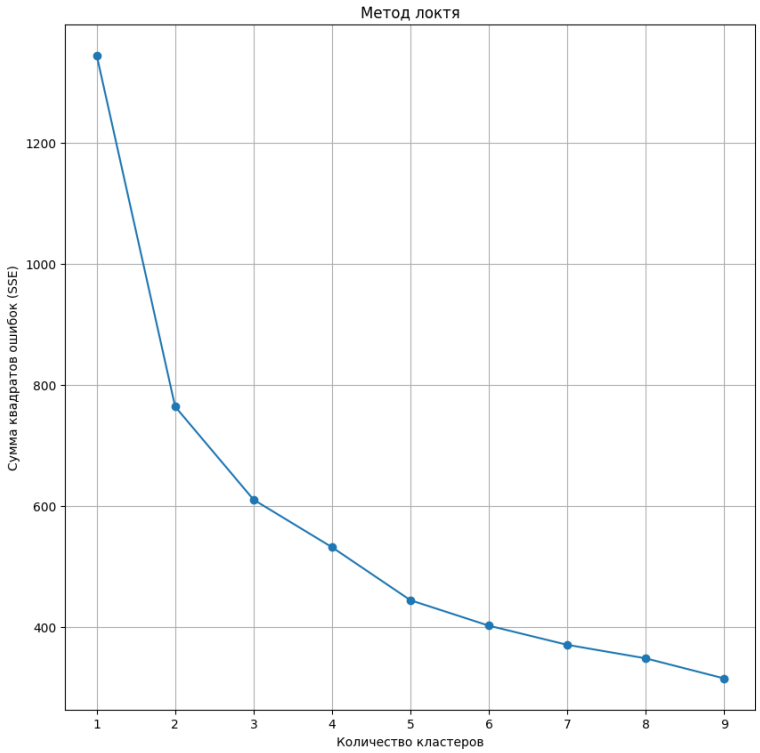

Рисунок 14 – Результат применения метода локтя

Из графика видно, что стоит выбрать 2 кластера, так как самое резкое изменение значения ошибки происходит при выборе 2 кластеров. Но 2 кластера недостаточно, поэтому были попытки создать количество кластеров от 4 до 7. После анализа полученных кластеров стало понятно, что данный алгоритм не подходит для кластеризации, потому что нельзя было выделить какие-то общие признаки внутри кластеров.

Алгоритм иерархической кластеризации с методом дальнего соседа [11] показал гораздо лучшие результаты. Метод ближнего соседа не был выбран, так как в данных нет многочисленного шума. На рисунке 15 показан код для кластеризации.

```
# complete -- метод дальнего соседа
agglomerative = AgglomerativeClustering(n_clusters=6, linkage='complete')
```

Рисунок 15 – Иерархическая кластеризация методом дальнего соседа

После кластеризации к исходным данным добавлено поле cluster. Дополненный информацией файл имеет название data_agglomerative.csv.

2.4 Анализ полученных кластеров

Для построения графиков выбрана цветовая схема от красного до зеленого для визуализации графиков, использующая такую последовательность цветов [26]: от '#FF0000' до '#00FF00', проходя через '#FF4000', '#FF8000', '#FFFF00' и '#80FF00'.

На рисунке 16 представлено, сколько образовательных организаций содержится в каждом кластере. Меньше всего объектов в 6 кластере, а больше всего в 4.

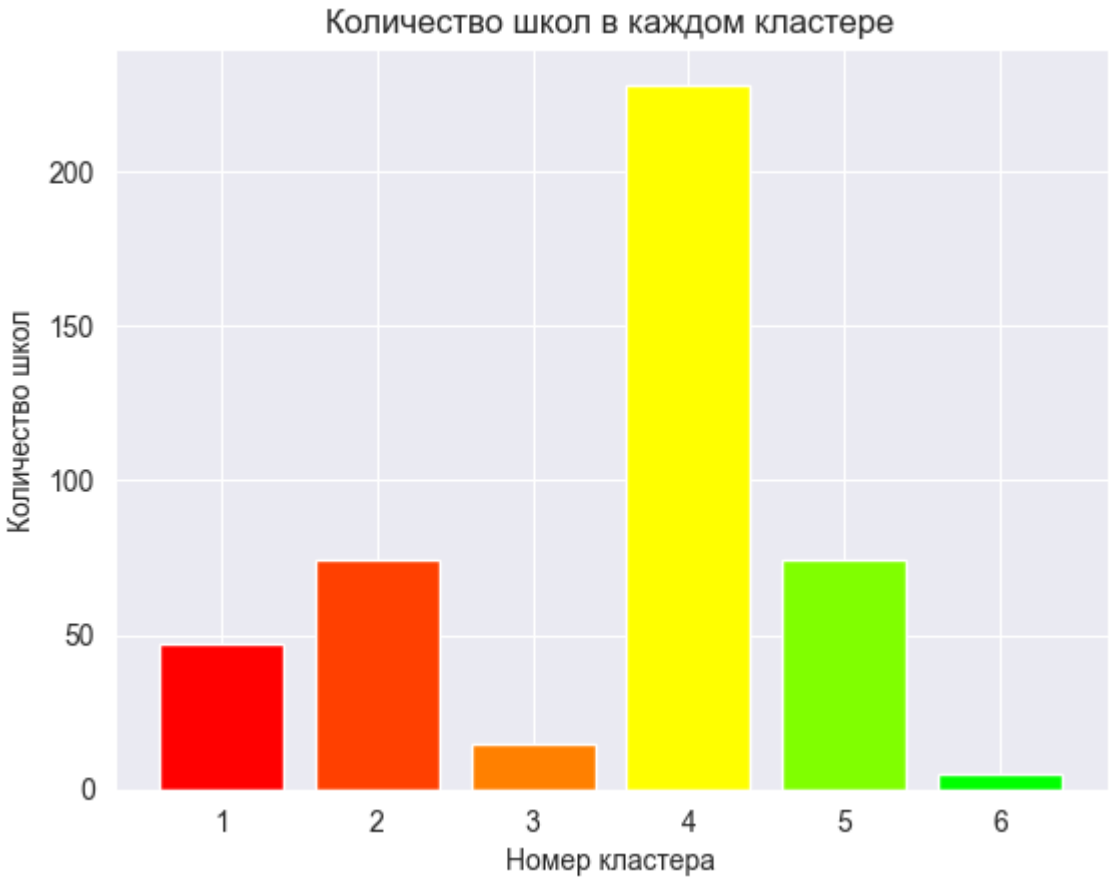

Рисунок 16 – Количество объектов в каждом кластере

Построены ящичковые диаграммы для всех полей, участвующих в кластеризации. После их просмотра можно сделать вывод, что данные хорошо сгруппированны, так как почти нет выбросов. Полученные ящичковые диаграммы изображены на рисунках 17-22.

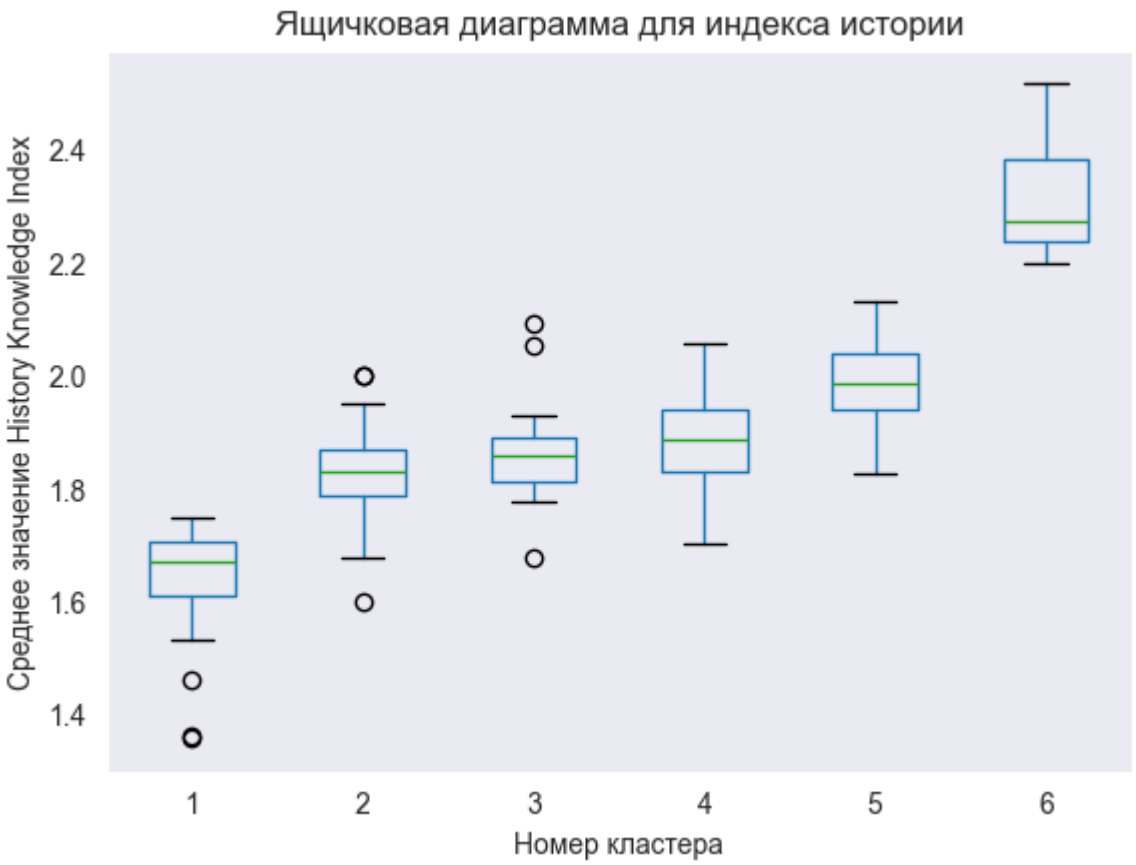

Рисунок 17 – Ящичковая диаграмма для индекса истории

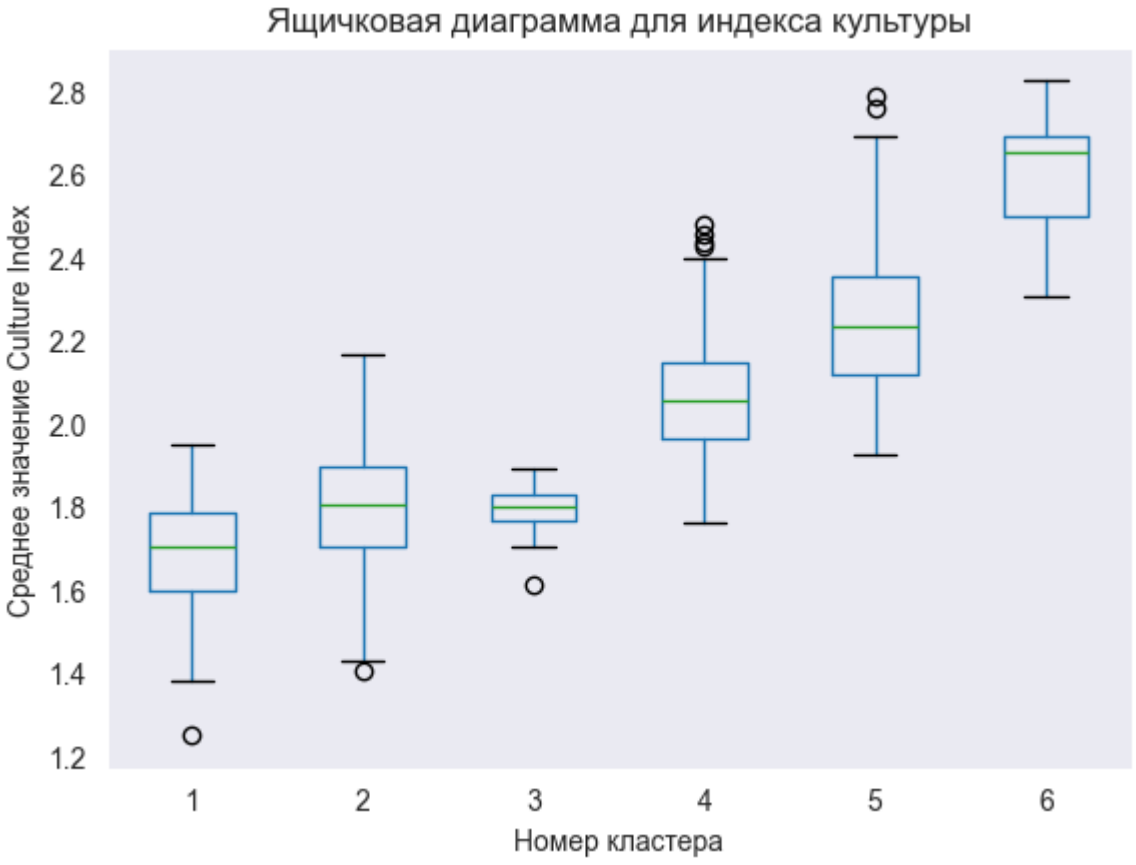

Рисунок 18 – Ящичковая диаграммадля индекса культуры

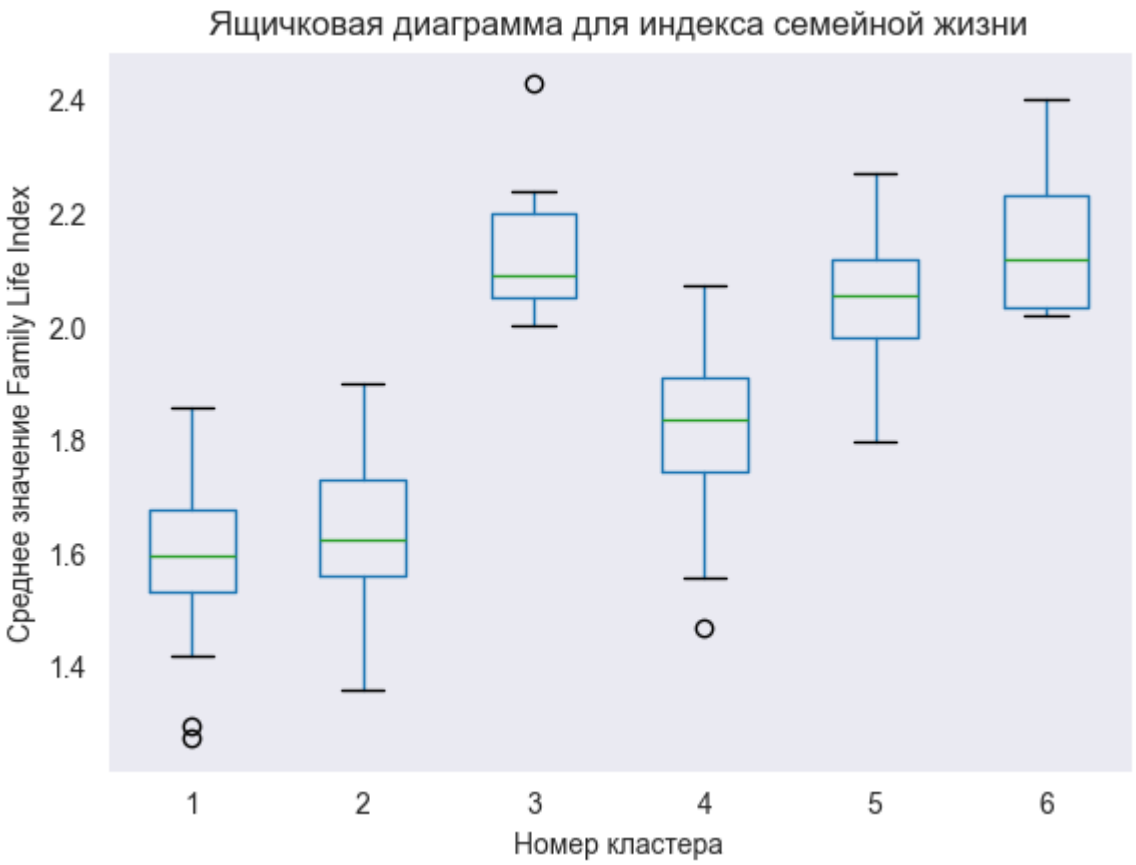

Рисунок 19 – Ящичковая диаграмма для индекса семейной жизни

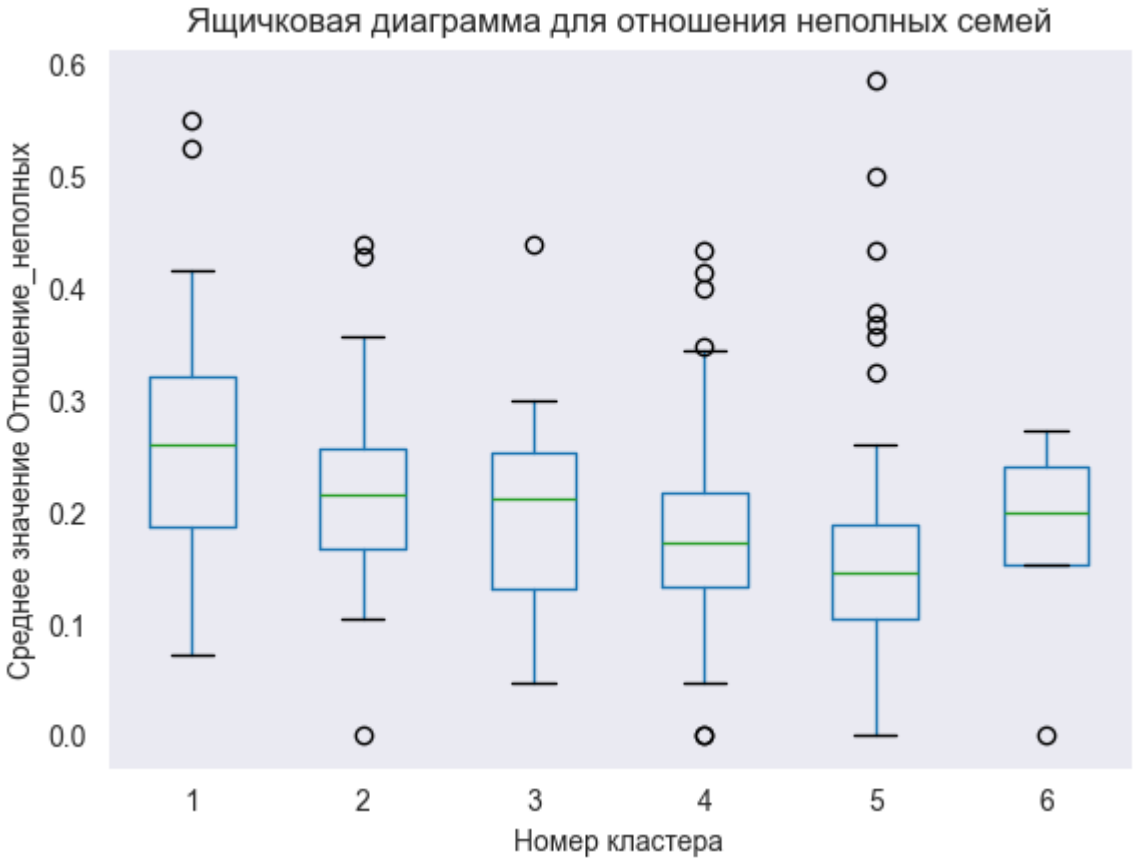

Рисунок 20 – Ящичковая диаграмма для Отношение_неполных

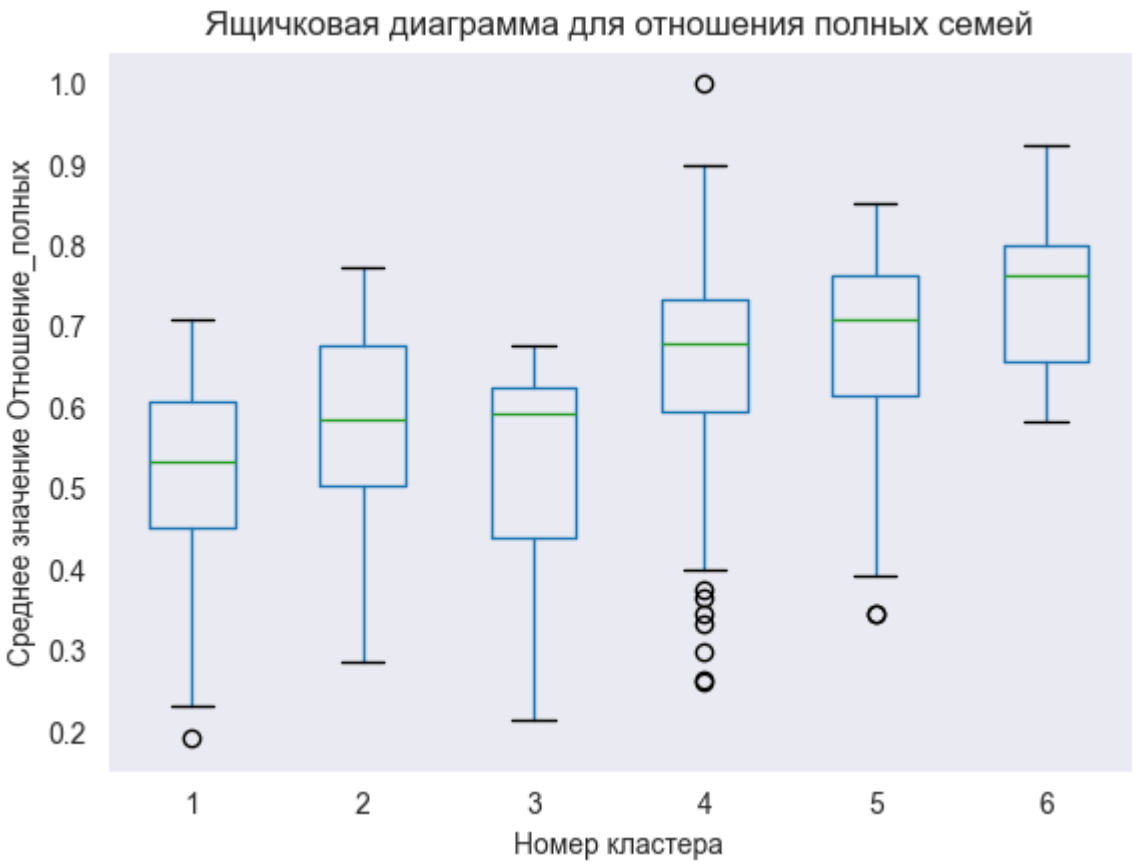

Рисунок 21 – Ящичковая диаграмма для Отношение_полных

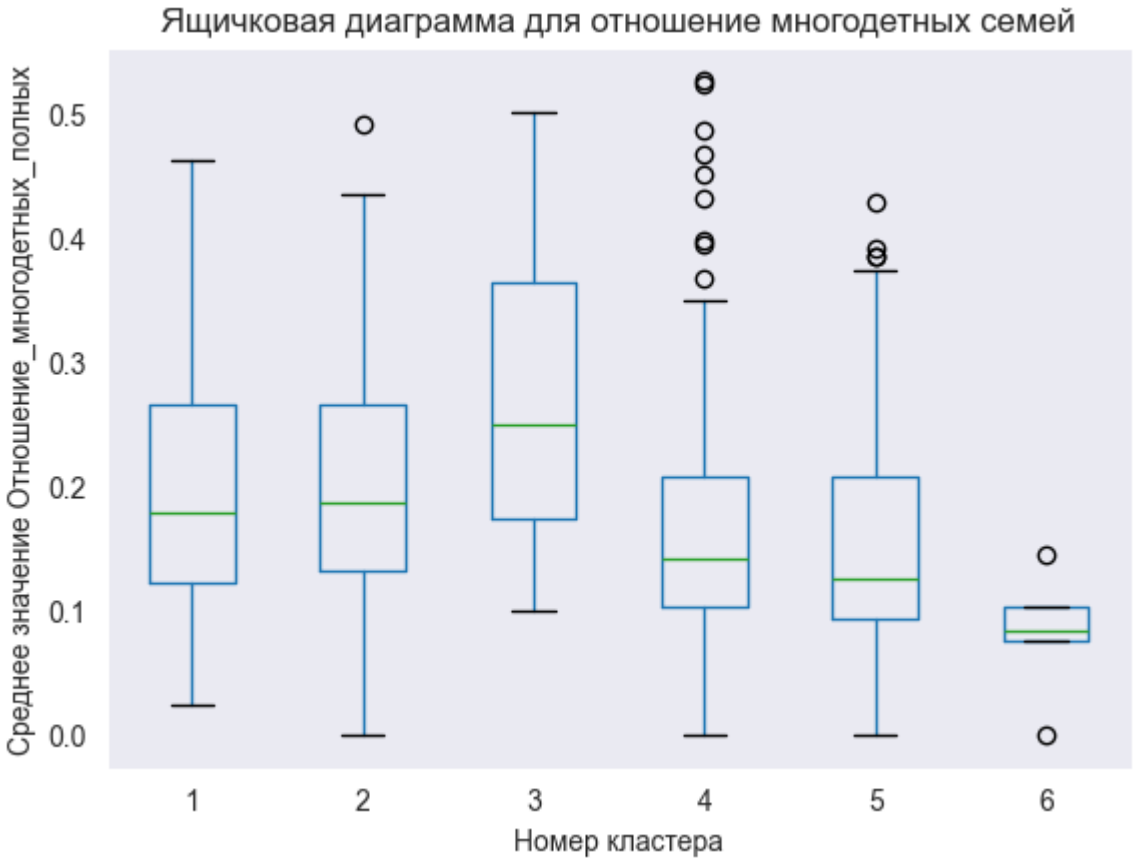

Рисунок 22 – Ящичковая диаграмма для Отношение_многодетных_полных

После просмотра графиков можно сделать первые выводы. Первый

кластер имеет самые низкие значения индексов, а также самое высокое значение отношение неполных семей и самое низкое отношение полных семей. Шестой кластер, наоборот, имеет самые высокие показатели индексов, а отношение полных семей принимает самое высокое значение среди всех кластеров. Из-за третьего кластера нарушается тенденция, что у кластеров с меньшим номером меньше значения индексов, чем у кластеров с большим номером. Высокое значение индекса семейной жизни у третьего кластера можно обосновать самым высоким показателем отношения многодетных полных семей в кластере.

Для дальнейшего анализа построены графики средних значений оценок по математике и русскому языку и первичному баллу за диагностическую работу в кластерах. Из рисунков 23-25 видно, что рассматриваемые значения возрастают с номером кластера. Разные кластера содержат разные значения этих полей, хотя эти поля не участвовали в кластеризации.

Рисунок 23 – Среднее значение оценки по математике в кластерах

Рисунок 24 – Среднее значение оценки по русскому языку в кластерах

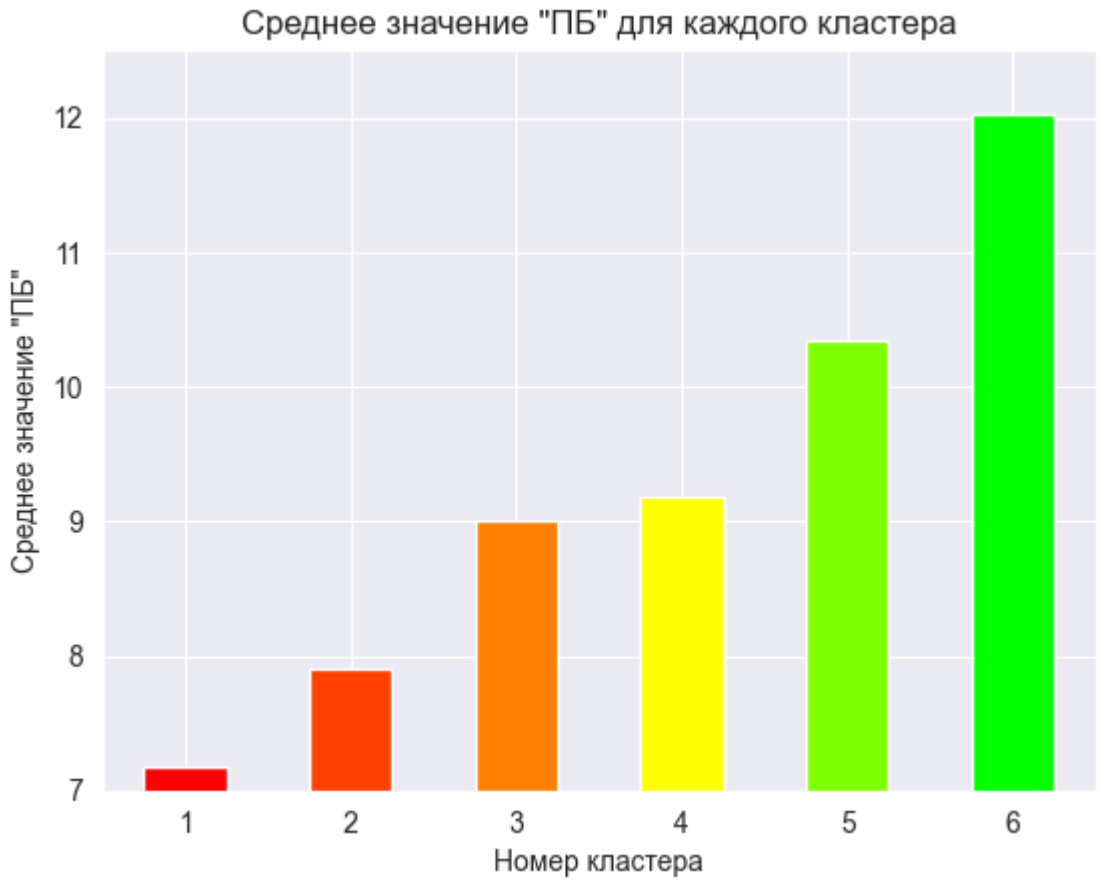

Рисунок 25 – Среднее значение оценки по первичному баллу в кластерах

Построен график средних зарплат школ из кластеров. Из рисунка 26 видно, что не смотря на самые высокие показатели индексов и первичного балла в кластере 6, заработная плата в нем самая низкая.

Рисунок 26 – Средняя заработная плата учителей в кластерах

На рисунках 27-29 с средними значениями индексов истории, культуры, семейной жизни по кластерам также наблюдается рост значений вместе с ростом номера кластера.

Рисунок 27 – Среднее значение индекса истории для каждого кластера

Рисунок 28 – Среднее значение индекса культуры для каждого кластера

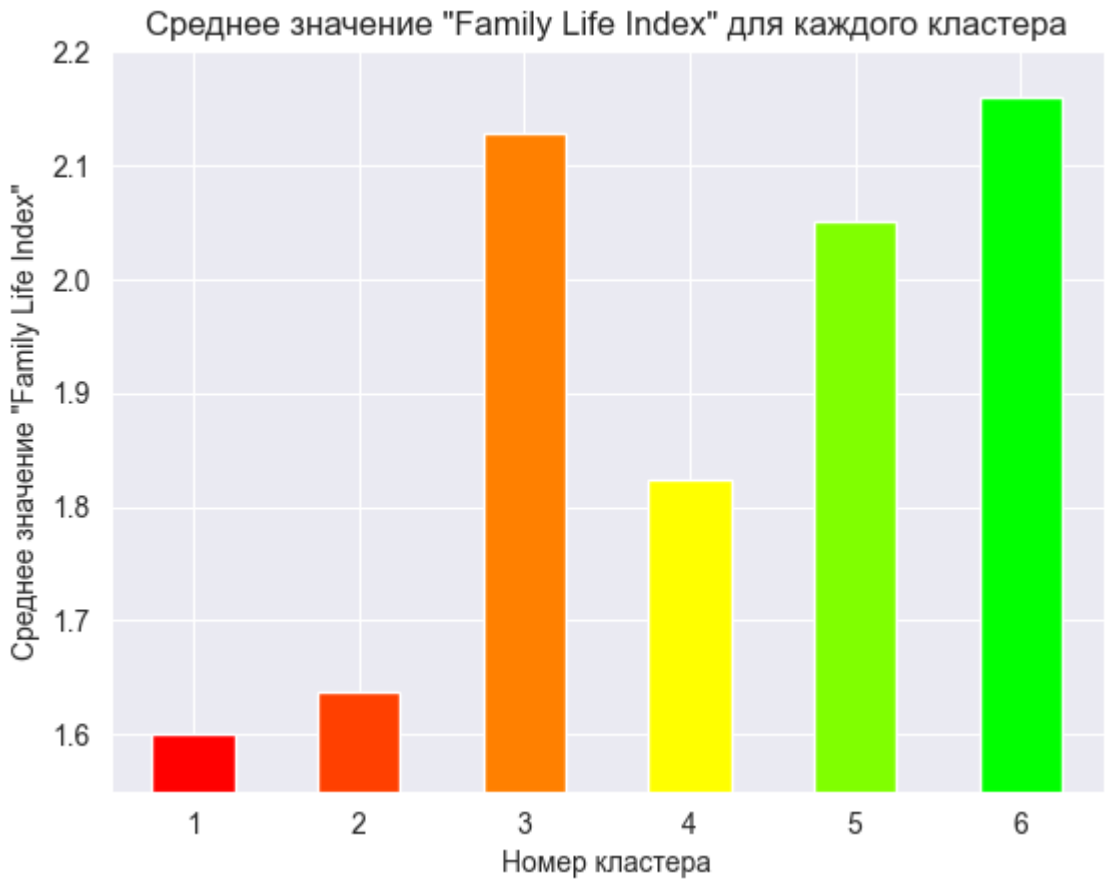

Рисунок 29 – Среднее значение индекса семейной жизни для каждого кластера

Из предыдущих выводов и рисунков 30-32 видно, что чем больше значение отношения полных семей, тем выше значение индексов и первичного балла за работу. И, наоборот, чем выше отношение неполных семей, тем ниже показатели.

Рисунок 30 – Среднее значение «Отношение_полных» в кластерах

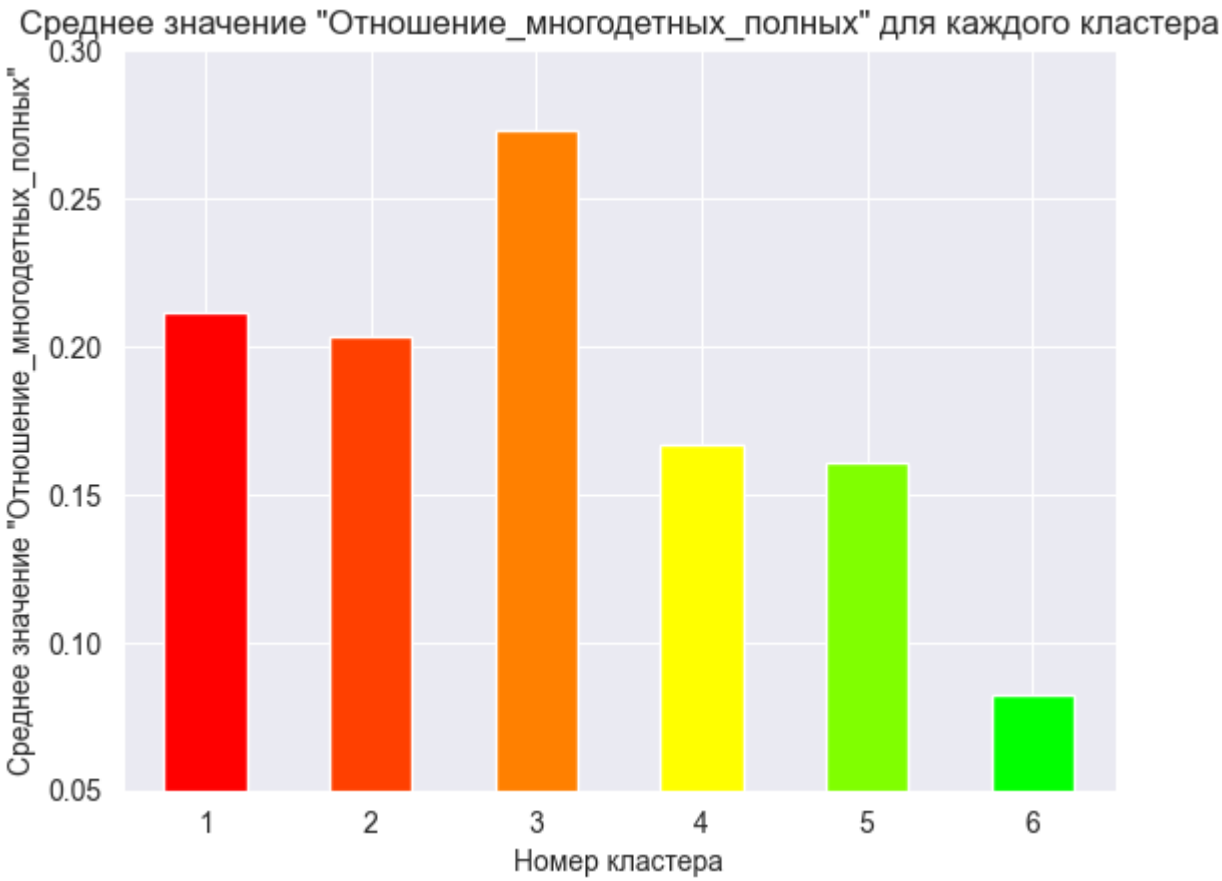

Рисунок 31 – Среднее значение «Отношение_многодетных_полных» в кластерах

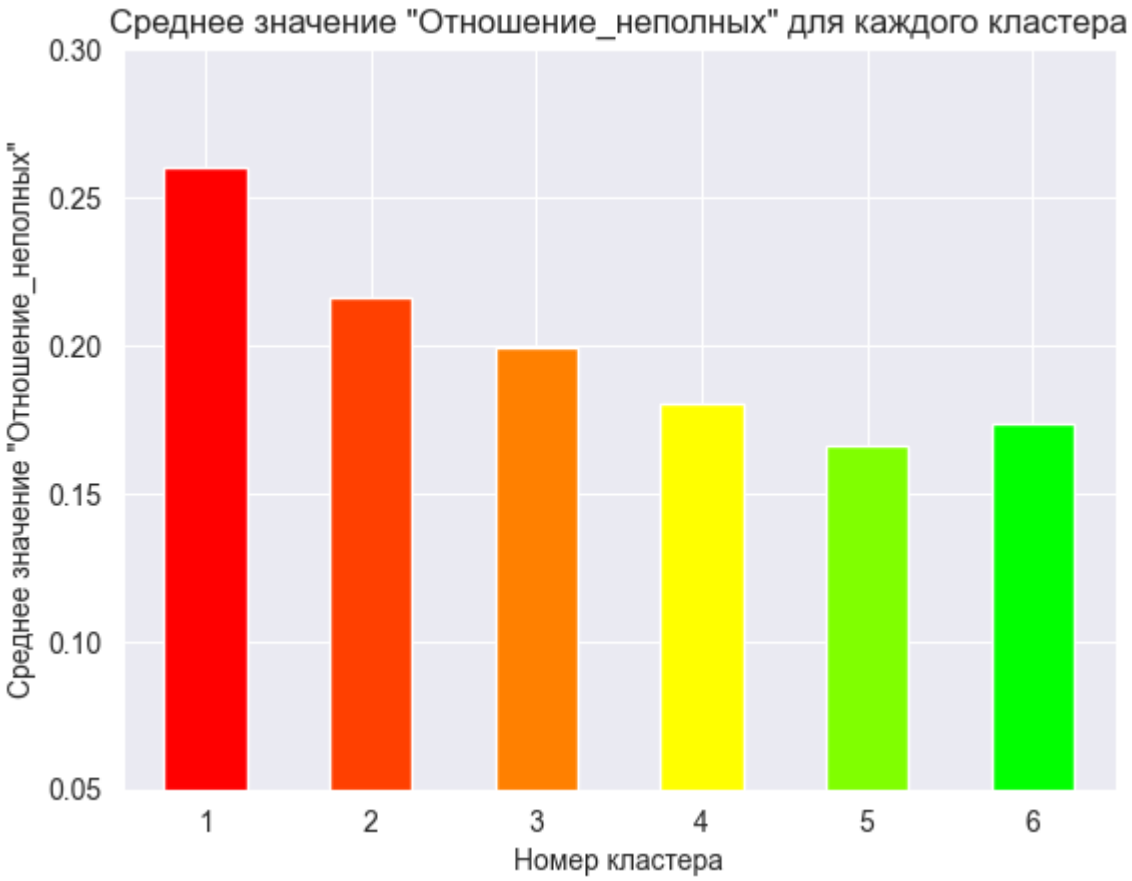

Рисунок 32 – Среднее значение «Отношение_неполных» в кластерах

Для просмотра того, как данные сгруппированны, постреены диаграммы рассеяния. Они показаны на рисунках 33-34.

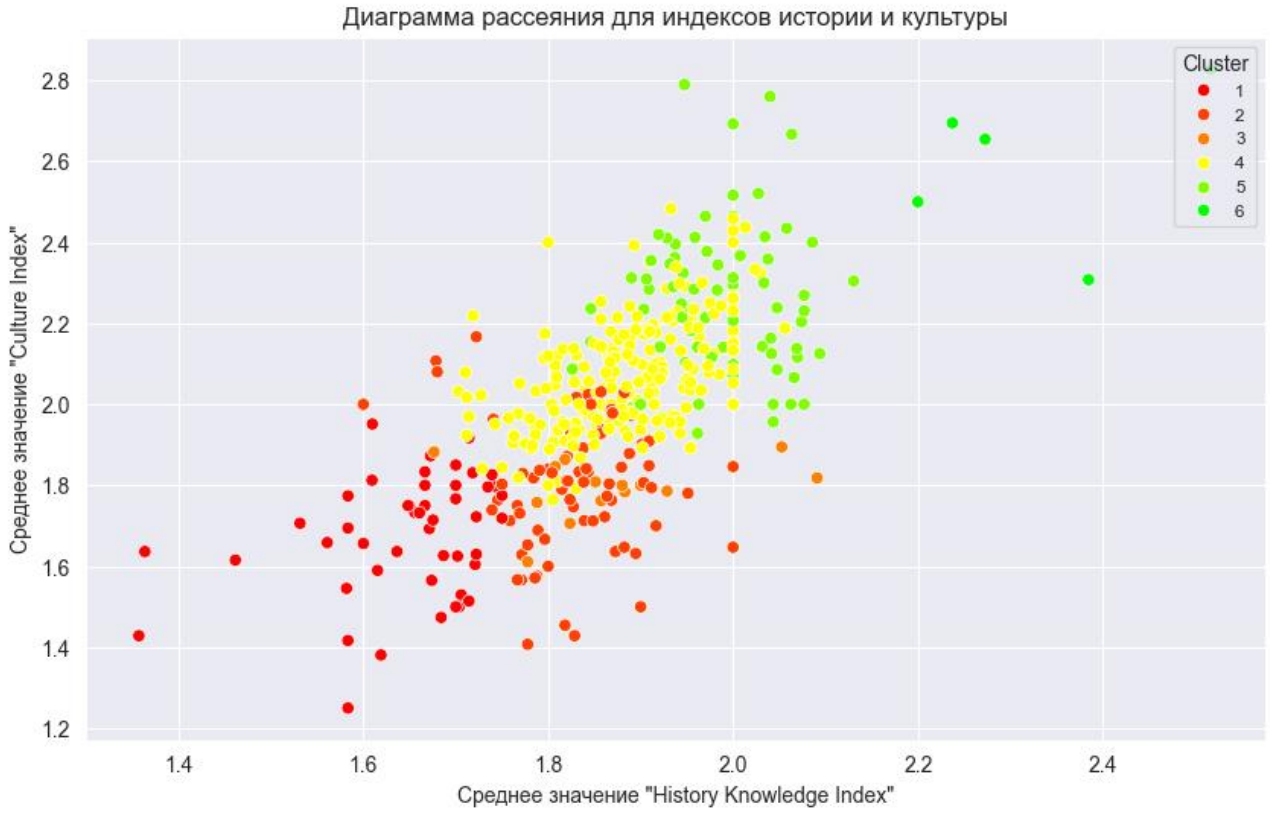

Рисунок 33 – Диаграмма рассеяния между индексами культуры и истории

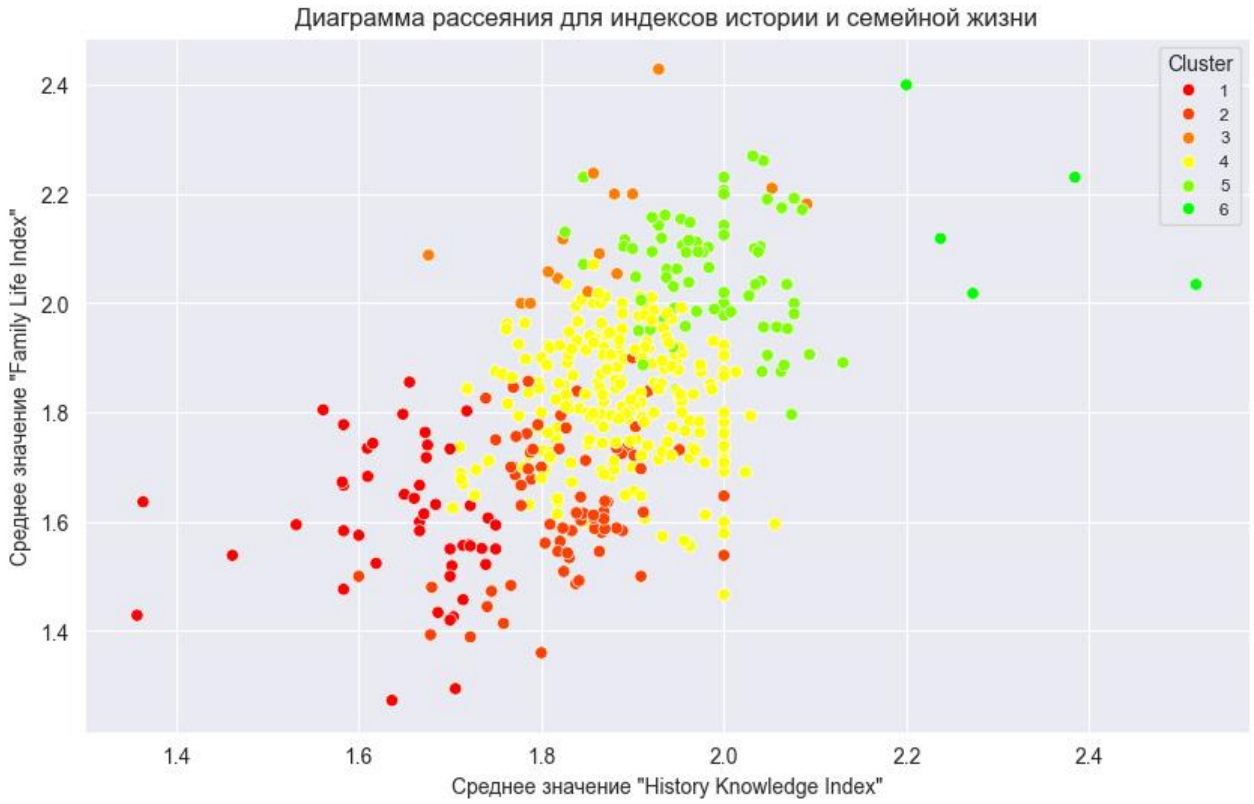

Рисунок 34 – Диаграмма рассеяния между индексами семейной жизни и истории

2.5 Тепловые карты кластеров

Для построения тепловых карт был скачан geojson файл с данными о России [18]. Он считывается в библиотеке geopandas, построение карт происходит с помощью библиотеки folium. Тепловые карты интерактивны, они сохраняются в формате html [30].

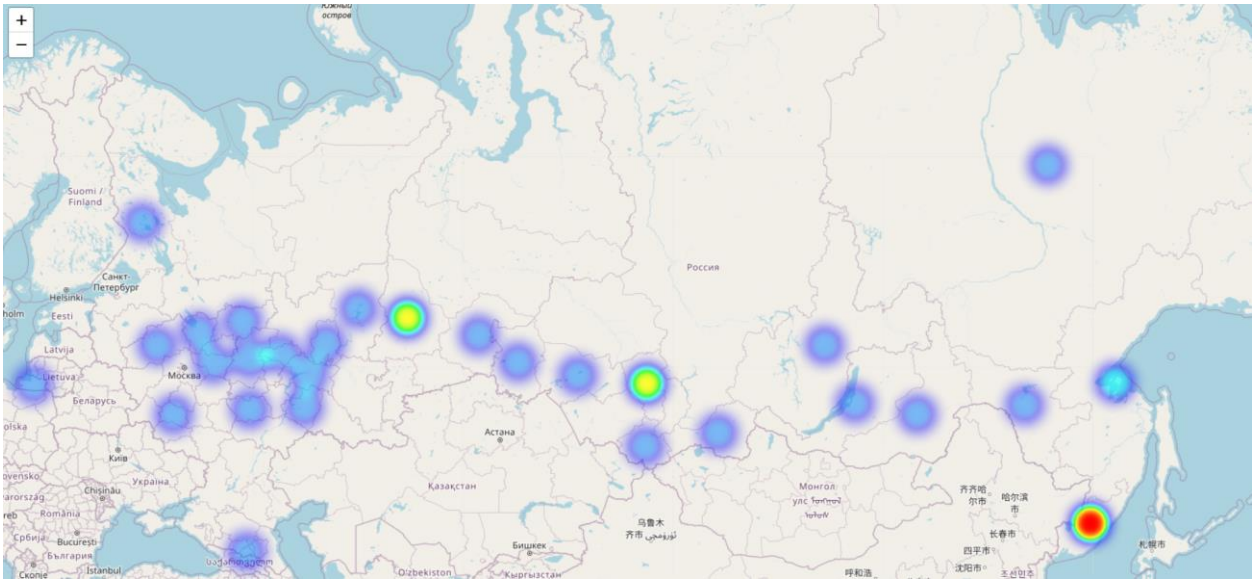

Рисунок 35 – Тепловая карта со школами из первого кластера

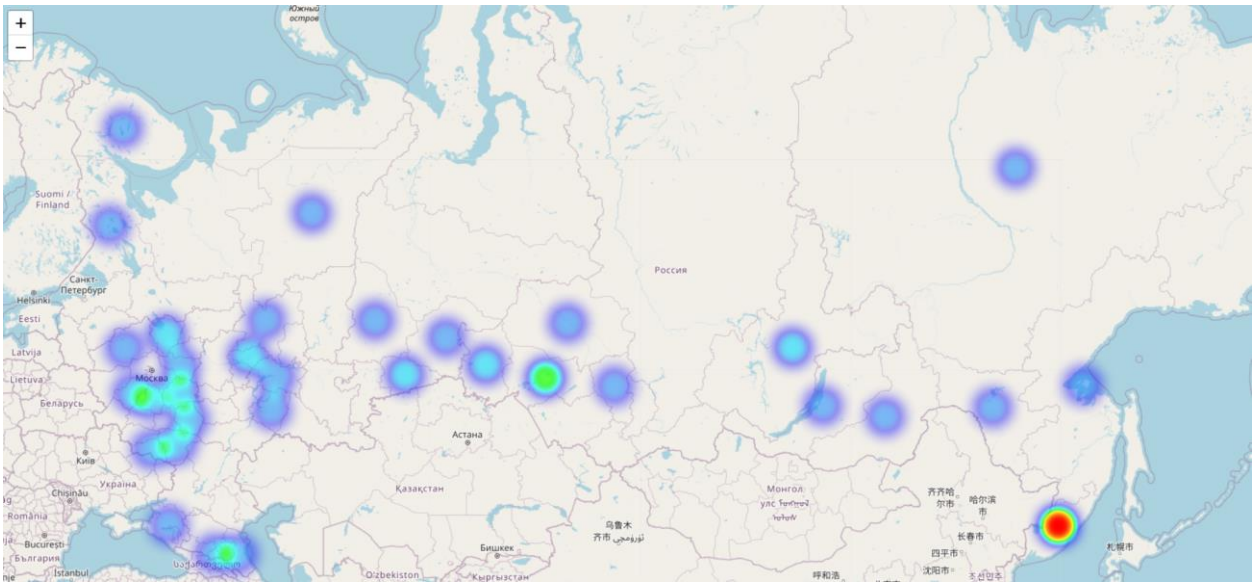

Рисунок 36 – Тепловая карта со школами из второго кластера

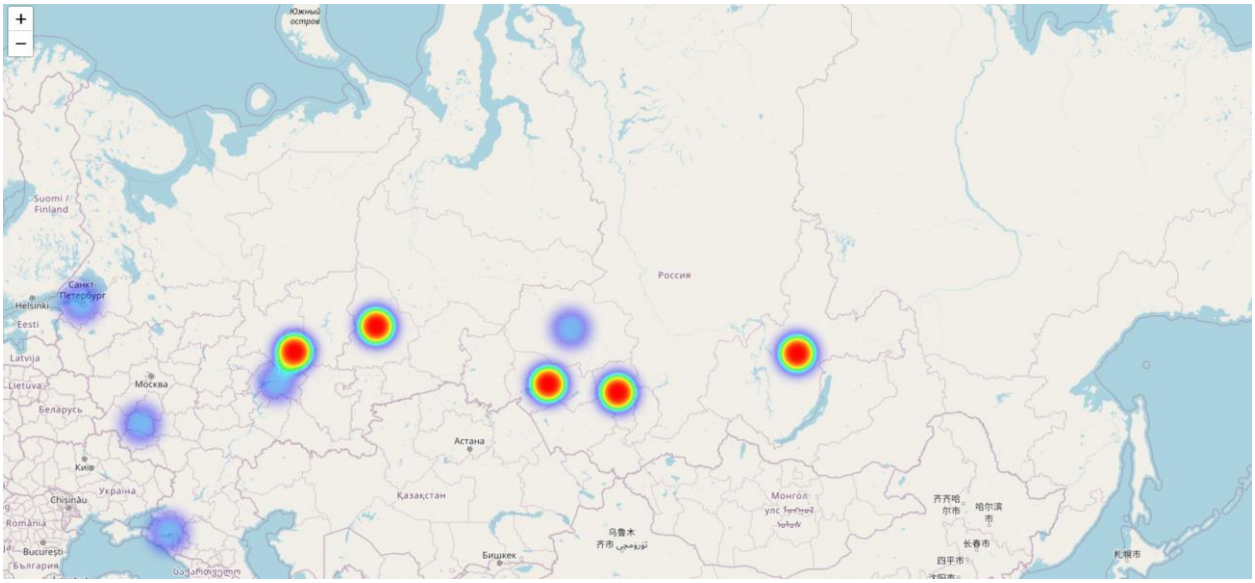

Рисунок 37 – Тепловая карта со школами из третьего кластера

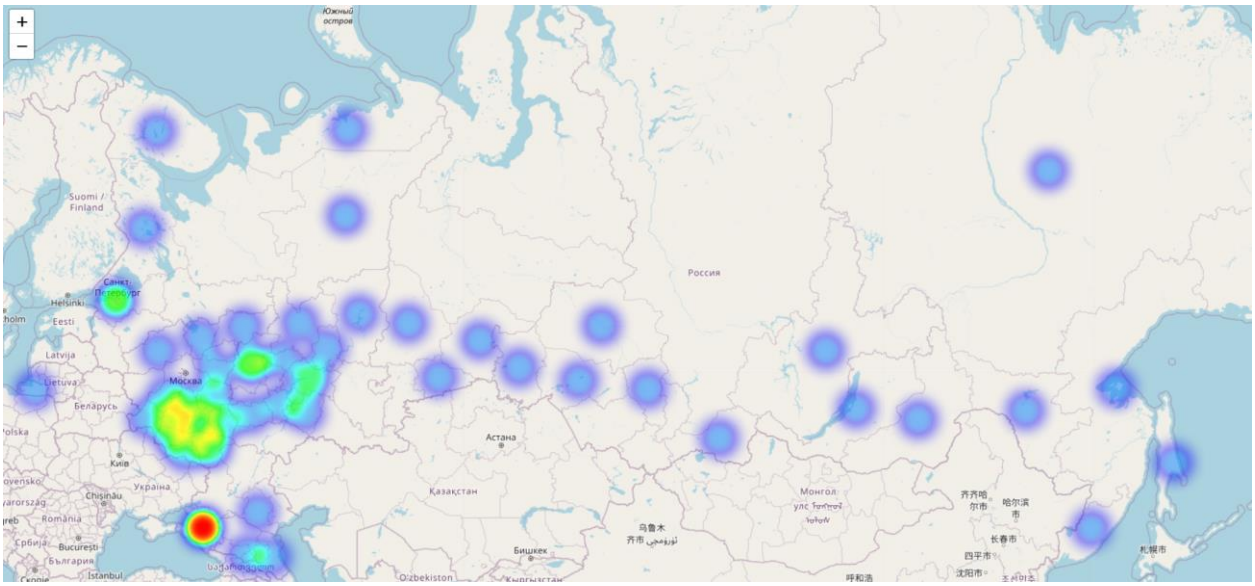

Рисунок 38 – Тепловая карта со школами из четвертого кластера

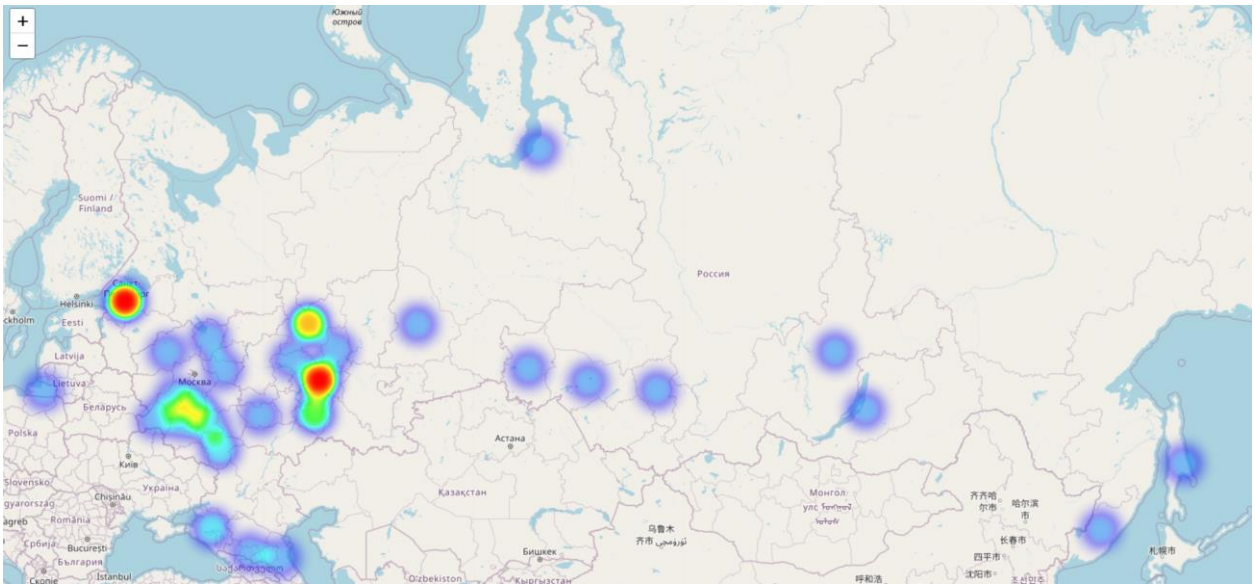

Рисунок 39 – Тепловая карта со школами из пятого кластера

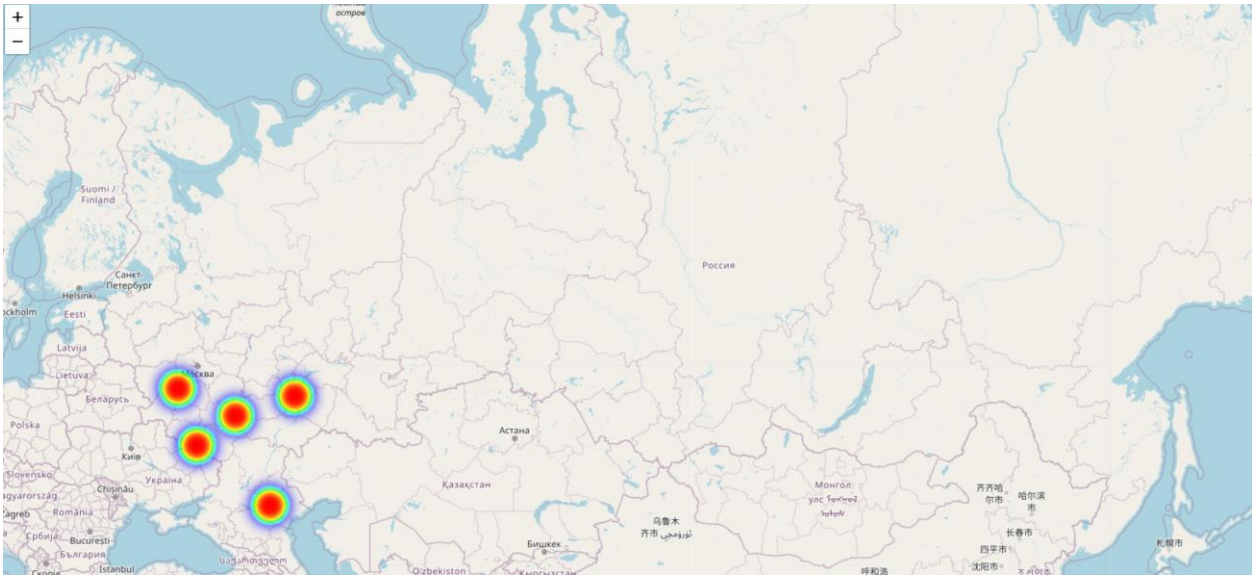

Рисунок 40 – Тепловая карта со школами из шестого кластера

3 РЕЗУЛЬТАТЫ РАБОТЫ

После просмотра графиков и их анализа можно описать полученные профили образовательных организации, обозначив номерами кластера:

1. Слабые школы

Школы с низкими показателями успеваемости, высоким процентом неполных семей и низким процентом полных семей. В основном расположены в Свердовском и Кемеровском округах и Приморском крае.

2. Школы с удовлетворительной успеваемостью

Школы со слабо-средними показателями. Процент неполных семей выше среднего. В основном расположены в Новосибирской области и Приморском крае.

3. Школы с направленностью в СПО.

Обычные школы с более низким индексом культуры, но с высоким уровнем семейной жизни. Имеют высокий процент многодетных семей. В основном расположены в Удмуртии, Свердловской, Новосибирской, Кемеровской, Иркутской областях – в промышленных зонах, здесь развито СПО [25], но особо нет объектов культуры.

4. Среднестатистические школы

Школы со средними показателями, распределены по всей карте России. Таких школ большинство.

5. Школы с высокими показателями

Школы с высокими показателями всех индексов. Состав семьи как у среднестатистических школ. В основном расположены в Северно-Западном, Центральном, Приволжском округах.

6. Резельентные школы

Резельентные школы [28] с самыми высокими показателями, ученики в основном из полных семей. Резельентность школ обусловлена тем, что они не находятся в центральных регионах, но показывают высокую успеваемость и учителя школ из этого профиля имеют заработную плату нижу, чем в школах из других профилей.

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы была проведена работа, ориентированная на решение актуальной задачи кластеризации образовательных организаций. На языке программирования Python разработаны специализированные модули, что позволило автоматизировать процесс группировки школ.

Для аналитического обзора полученных профилей был применен комплексный подход, включающий в себя исследование основных параметров и особенностей каждого выделенного кластера. Описание составленных профилей демонстрирует разнообразие образовательной среды и способствует повышению понимания структурных особенностей школьного образования в исследуемых регионах.

Одним из наиболее информативных инструментов визуализации данных стали тепловые карты. Они визуально демонстрируют распределение школ по кластерам в пространственном измерении, что не только наглядно представляет исходные данные, но и может служить основой для дальнейших исследований и разработки рекомендаций по оптимизации работы образовательных учреждений и планированию новых.

Важной частью работы является и оценка качества проведенной кластеризации, что позволяет убедиться в эффективности выбранных методов и корректности интерпретации результатов. Возможности по улучшению методологии могут включать в себя глубокий анализ и внедрение различных метрик для оценки качества кластеров, а также использование дополнительных данных и параметров.

Таким образом, работа несет в себе значительный аналитический потенциал и может быть использована как в органах управления образованием, так и при разработке стратегий развития образовательной инфраструктуры на муниципальном или региональном уровне.

Проект имеет потенциал для дальнейшего расширения. Расширение диапазона данных, использованных при кластеризации, может способствовать улучшению детализации и разнообразия получаемых профилей образовательных учреждений. Кроме того, с применением технологий искусственного интеллекта можно не только группировать школы по кластерам, но и осуществлять более глубокое описание и анализ индивидуальных характеристик каждого учебного заведения.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Python3 документация. – URL: https://docs.python.org/3/ (дата обращения: 11.02.2024).

2. Pandas документация. – URL: https://pandas.pydata.org/docs/ (дата обращения: 15.02.2024).

3. Matplotlib документация. – URL: https://matplotlib.org/ (дата обращения 15.02.2024).

4. Seaborn документация. – URL: https://seaborn.pydata.org/ (дата обращения: 15.02.2024).

5. Sklearn документация. – URL: https://scikit-learn.org/ (дата обращения 01.03.2024).

6. Geopandas документация. – URL: https://geopandas.org/en/stable/gallery/plotting_with_folium.html (дата обращения 07.03.2024).

7. Geojson документация. – URL: https://doc.arcgis.com/ru/arcgis-online/reference/geojson.htm (дата обращения 20.03.2024).

8. Folium документация. – URL: https://python-visualization.github.io/folium/latest/ (дата обращения 07.03.2024).

9. Sklearn StandartScaler документация. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (дата обращения 10.03.2024).

10. Sklearn KMeans документация. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html (дата обращения 10.03.2024).

11. Sklearn Agglomerative Clustering документация. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html (дата обращения 10.03.2024).

12. Датасет с зарплатами учителей в России за 2020 год. – URL : https://rosstat.gov.ru/storage/mediabank/n3VTgP5l/04-20-03.xlsx (дата обращения: 15.03.2024).

13. Создание Северно-Кавказского федерального округа. – URL: https://riadagestan.ru/news/politics/sem_let_nazad_byl_sozdan_severo_kavkazskiy_federalnyy_okr (дата обращения 15.03.2024).

14. Pandas merge документация. – URL: https://pandas.pydata.org/docs/reference/api/pandas.merge.html (дата обращения: 25.03.2024).

15. Культурная столица России. – URL:

http://migrantinfo.kmormp.gov.spb.ru/informaciya-o-sankt-peterburge/sankt-peterburg-kulturnaya-stolica/ (дата обращения 16.03.2024).

16. TSNE документация. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html (дата обращения 27.03.2024).

17. Метод локтя. – URL: https://www.dmitrymakarov.ru/intro/clustering-16/ (дата обращения 27.03.2024).

18. Russia geojson. – URL: https://github.com/timurkanaz/Russia_geojson_OSM/tree/master/GeoJson's (дата обращения 27.03.2024).

19. О программе НИКО. – URL: https://fioco.ru/ru/osoko/niko/ (дата обращения 16.03.2024).

20. Отчет НИКО за 2020 год. – URL: https://fioco.ru/Media/Default/Documents/NIKO/%D0%9E%D1%82%D1%87%D0%B5%D1%82%20%D0%9D%D0%98%D0%9A%D0%9E%202020.pdf (дата обращения 20.03.2024).

21. Ящичковая диаграмма. – URL: https://datavizcatalogue.com/RU/metody/diagramma_razmaha.html (дата обращения 19.03.2024).

22. Диаграмма рассеяния. – URL: https://datavizcatalogue.com/RU/metody/diagramma_rassejanija.html (дата обращения 20.03.2024).

23. Агрегирующий функции pandas. – URL: https://teletype.in/@dt_analytic/jXHX36Vd_cC (дата обращения 26.03.2024).

24. Медианное значение. – URL: https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%B4%D0%B8%D0%B0%D0%BD%D0%B0_(%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0) (дата обращения 02.03.2024).

25. Среднее профессиональное образование в России. – URL: https://ru.wikipedia.org/wiki/%D0%A1%D1%80%D0%B5%D0%B4%D0%BD%D0%B5%D0%B5_%D0%BF%D1%80%D0%BE%D1%84%D0%B5%D1%81%D1%81%D0%B8%D0%BE%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%80%D0%B0%D0%B7%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%B2_%D0%A0%D0%BE%D1%81%D1%81%D0%B8%D0%B8 (дата обращения 06.04.2024).

26. Таблица HTML цветов. – URL: https://colorscheme.ru/html-colors.html (дата обращения 01.04.2024).

27. Обзор алгоритмов кластеризации. – URL: https://habr.com/ru/articles/101338/ (дата обращения 05.03.2024).

28. Резильентные школы. – URL: https://trends.rbc.ru/trends/education/5f772ce89a79471dab5eef06 (дата обращения

03.04.2024).

29. Группировка данных в pandas. – URL: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.groupby.html (дата обращения 12.03.2024).

30. HTML документация. – URL: https://developer.mozilla.org/en-US/docs/Web/HTML (дата обращения 15.04.2024).

ПРИЛОЖЕНИЕ А

Исходный код

На рисунке А.1 изображен QR-код со ссылкой на GitHub репозиторий с исходным кодом разработанного программного продукта.

Рисунок А.1 – QR-код на репозиторий