­­­­­­­Московский Авиационный Институт

(Национальный исследовательский университет)

Выпускная квалификационная работа.

Тема: Корреляционный анализ результатов оценочных процедур в образовательных организациях.

Выполнил студент группы 80-408Б: Дубов А.В.

Научный руководитель: к.ф.-м.н., доц. Левинская М.А.

Москва

2022 год

# РЕФЕРАТ

Выпускная квалификационная работа бакалавра состоит из 33 страниц, 23 рисунков, 7 таблиц, 10 использованных источников, 2 приложений.

ОЦЕНКА КАЧЕСТВА ОБРАЗОВАНИЯ, АНАЛИЗ ДАННЫХ, SQL, PYTHON, PANDAS, СБОР ДАННЫХ, СКРЕЙПИНГ, ПАРСИНГ, ОБРАЗОВАТЕЛЬНЫЕ ОРГАНИЗАЦИИ, КОНТЕКСТНЫЕ ДАННЫЕ, УСПЕВАЕМОСТЬ, ШКОЛЫ

Цель работы – анализ корреляций результатов различных оценочных процедур в образовательных организациях с различными контекстными данными.

В результате предполагается получить данные необходимые для кластеризации общеобразовательных организаций.

# СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 4](#_Toc105618031)

[1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 7](#_Toc105618032)

[1.1 Постановка задачи и проблема, которую решает ВКР 7](#_Toc105618033)

[1.2 Выбор стека технологий 9](#_Toc105618034)

[1.2.1 Парсинг и скрейпинг 9](#_Toc105618035)

[1.2.2 Плюсы и минусы использования библиотеки beautifulsoup4 10](#_Toc105618036)

[1.2.3 Искусственный Интеллект в парсинге 11](#_Toc105618037)

[1.2.4 Обработка больших объемов данных 11](#_Toc105618038)

[1.2.5 Выбор среды для парсинга 13](#_Toc105618039)

[1.2.6 Выбор среды обработки данных 13](#_Toc105618040)

[1.2.7 Расчет корреляции 14](#_Toc105618041)

[2 ПРАКТИЧЕСКАЯ ЧАСТЬ 15](#_Toc105618042)

[2.1 Исходные данные 15](#_Toc105618043)

[2.2 Описание/представление результатов 20](#_Toc105618044)

[2.2.1 Парсинг 20](#_Toc105618045)

[2.2.2 Обработка данных 22](#_Toc105618046)

[2.2.3 Результаты расчета корреляции 23](#_Toc105618047)

[ЗАКЛЮЧЕНИЕ 25](#_Toc105618048)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 26](#_Toc105618049)

[ПРИЛОЖЕНИЕ А Код программы 28](#_Toc105618050)

[ПРИЛОЖЕНИЕ Б Дополнительные данные 33](#_Toc105618051)

# ВВЕДЕНИЕ

Общеобразовательные организации начали появляться в России с 1919 года и имеют огромное значение для государства и общества. Уровень образования в ОО влияет на показатели экономики, производительность труда и другие аспекты, которые влияют на качество жизни населения. Из-за этого важно иметь представление о качестве образования в ОО.

Качество образования состоит из:

– Качества результатов подготовки обучающихся и степени достижения планируемых результатов образовательной программы.

– Качества содержания образовательных программ.

– Качества различных условий (количество и уровень компетентности работников, вариативность психологического и педагогического процессов, санитарно-гигиенические условия, медицина, организация питания, общественное управление).

Оценка качества образования может проводится разнообразными способами, такими как:

– на государственном уровне – соответствием результатов образования требованиям государственных образовательных стандартов;

– на уровне образовательного учреждения качество образования определяется соответствием его результатов образовательному стандарту, действующему в пределах данного учреждения и зафиксированному в соответствующих образовательных программах.

Для оценки качества образования проводят экзамены, ставят отметки; собирают данные с помощью социологических опросов.

Также представлен подход к оценке результатов деятельности образовательных организаций, позволяющий учесть их социальную специфику. Он подразумевает использование контекстных данных в системе оценки качества образования. Идея заключается в том, что при наличии необходимых данных можно выявить устойчивые соотношения между учебными достижениями учащихся и контекстными данными (информация о том, где находится школа, культурный контекст, объем ВВП и т.п.).

Актуальность исследованиясвязана с постоянной необходимостью усовершенствования процесса обучения. Нахождение закономерностей, отражающихся на качестве образования, позволит обратить внимание на недостатки и даст возможность исправить их.

Цельисследования – анализ корреляций результатов различных оценочных процедур в образовательных организациях с различными контекстными данными.

Объект исследования – успеваемость в образовательных организациях.

Предметом исследования является влияние контекстных факторов (геолокация, культурный контекст, транспортная доступность) на успеваемость в ОО.

Для достижения поставленной цели необходимо было решить следующие задачи:

* изучить документацию для обработки большого объема данных;
* выделить контекстные данные, которые влияют на успеваемость;
* создать dataframe с идентификаторами школ и их контекстными данными, для дальнейшего использования в кластеризации школ.

Для решения поставленных задач были использованы следующие методы исследования:

* изучение и анализ литературы по теме исследования;
* парсинг и веб-скрейпинг для получения данных с открытых сайтов;
* очистка данных;
* анализ и сортировка данных;
* вычисление корреляций между различными факторами.

База исследования: ВПР 2020-2021 гг.

ВПР – итоговые контрольные работы, проводимые по отдельным учебным предметам для оценки уровня подготовки школьников, которые начали проводиться с 2015 года. Это не аналог государственной итоговой аттестации, так как они проводятся на региональном или школьном уровне, а результаты ВПР не влияют на оценку учащегося в самом учебном заведении.

Данные, полученные с ВКР, можно использовать для оценки качества образования и принятия мер по устранению неполадок в системе образования.  
Также ВКР позволяет оценить качество и эффективность предпринятых мер.

Практическая значимость исследования заключается в нахождении зависимости успеваемости образовательных организаций от контекстных факторов и улучшению качества образования (в виде предсказывания уязвимых областей и проведения в них совершенствования), с помощью полученных зависимостей.

Результаты работы используются для дальнейшей кластеризации ОО по РФ и прогнозирования результатов ОО в следующем учебном году в рамках группового проекта ВКР.

# ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

## 1.1 Постановка задачи и проблема, которую решает ВКР

Машинное обучение хорошо подходит для анализа данных и оценки эффективности кластеров, однако для их создания необходимо найти, и очистить данные.

В работе решается проблема сбора данных, влияющих на качество образования, а также их последующую очистку и анализ. Предварительная обработка данных необходима для того, чтобы данные можно было использовать в дальнейшем для кластеризации в ином случае дальнейший анализ может быть невозможен из-за того, что алгоритмы, производящие аналитику, не смогут работать или результаты их работы будут некорректными. Этот шаг, включает в себя преобразование необработанных данных таким образом, чтобы проблемы, связанные с неполнотой, непоследовательностью и / или отсутствием надлежащего представления тенденций, решались таким образом, чтобы получить набор данных в читаемом формате.

Качество данных зависит от:

* Точности. Точность относится к тому, насколько хорошо записанная информация отражает реальное событие или объект. Неточность данных может быть вызвана неисправными приборами сбора данных или компьютерными ошибками, целенаправленной подачей неправильных значений данных пользователями, ошибками в передаче данных, несоответствиями в соглашениях об именах и форматах ввода и т. д.
* Полноты. Данные считаются “полными”, когда присутствуют все обязательные или необходимые функции. Неполнота данных может возникнуть из-за отсутствия необходимой информации, неисправностей оборудования во время сбора данных, непреднамеренного удаления или неспособности записать историю или изменения.
* Согласованности. Если одна и та же информация, хранящаяся и используемая в нескольких экземплярах, совпадает с несоответствиями форматирования или без них между различными источниками или хранилищами данных, то данные являются согласованными. Он количественно выражается как процент значений, которые совпадают между различными сохраненными экземплярами.
* Своевременности. Своевременность также влияет на качество данных, поскольку данные имеют ценность только тогда, когда они доступны при необходимости. Если данные устарели или исправления были внесены после оценки или анализа набора данных, это влияет на качество данных.
* Правдоподобия. Правдоподобность описывает доверие пользователей к данным. Если в какой-либо момент было обнаружено, что данные изобилуют ошибками и несоответствиями, его пользователи, скорее всего, будут иметь оговорки, когда дело доходит до использования этих данных в будущем. Поскольку это препятствует эффективному использованию данных по назначению, правдоподобность также является фактором, определяющим качество данных.
* Интерпретируемости. Интерпретируемость данных определяет, насколько легко понять информацию, присутствующую в наборе данных, и извлечь из нее смысл. Доступность методологий сбора и обработки статистических данных для пользователей может повлиять на интерпретируемость набора данных.

Хотя шесть измерений, описанных для оценки качества данных в этом разделе, ни в коем случае не являются исчерпывающим списком, они должны служить для ознакомления вас с многомерностью качества данных. Цель определения измерений качества данных остается прежней - убедиться, что данные могут удовлетворять требованиям их предполагаемого использования. Поэтому не страшно, если найдутся совершенно разные наборы измерений качества данных при обращении к различным источникам.

## Выбор стека технологий

### 1.2.1 Парсинг и скрейпинг

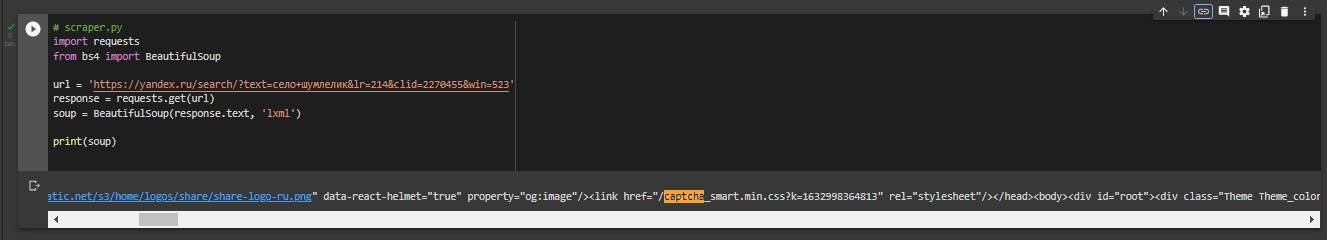
Сегодня наша жизнь напрямую связана с Интернетом и веб-сайтами: практически любая текстовая информация, которая нам может понадобиться, доступна онлайн. Рассмотрим, как собирать конкретную информацию с веб-сайтов.  
 HTML передает веб-браузерам, когда, где и какой элемент показывать на html-странице. HTML структура веб-страниц удобна для захвата информации.

Для того чтобы найти и упорядочить большое количество данных (например, соотнести села и города регионам) можно использовать веб скрейпинг открытых сайтов.

Найти необходимую информацию с помощью веб-сайтов можно либо через готовый API, либо использую парсинг. Веб-парсер это программа или скрипт, который используется для загрузки контента веб-страниц (обычно текста отформатированного на HTML) и извлечения из него данных.

Проблема в самостоятельном парсинге страниц заключается в том, что многие поисковые системы и сайты стараются его блокировать (например, Яндекс после нескольких запросов просит ввести капчу).

На рисунке 1.1 представлена проблема парсинга.



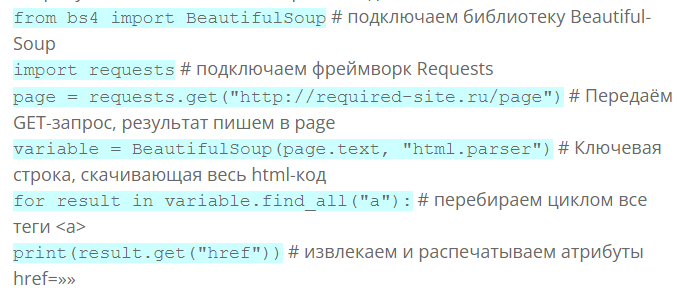
1. – Проблема парсинга

### Плюсы и минусы использования библиотеки beautifulsoup4

Плюсы:

Он прост в изучении. В частности, для получения ссылок со страницы потребуется всего несколько строчек кода.

На рисунке 1.2 представлен код получения ссылок со страницы.



1. – Код получения ссылок со страницы

Он имеет отличную подробную документацию, с помощью которой можно легко изучить функционал библиотеки.

Он обладает внушительной технической поддержкой.

Минусы:

Нужна установка фреймворка для отсылки запроса на веб-портал, потому что Beautiful-Soup не способен отправить его. Для выхода из этой ситуации необходимо воспользоваться одним из распространённых фреймворков: Requests, Lxml или Urlib2. Данные библиотеки сделают запрос к сайту. Наиболее часто применяется Requests, Lxml. После отправки информации HTML и XML на персональный компьютер желательно использовать программу для анализа скаченных данных.

Beautiful-Soup сравнительно медленно решает отдельную задачу, но за счёт использования технологии многопоточности легко обходит данную проблему.

### Искусственный Интеллект в парсинге

При использовании парсинга ИИ можно применять для задачи перевода органических нефильтрованных данных в машиночитаемый формат или машинное зрение для прохождения капчи. В том числе парсер может использовать синтаксический или лексический анализатор данных. Как правило, такие анализаторы построены с применением ИИ, поскольку иначе такую задачу решить трудно.

### Обработка больших объемов данных

Для обработки больших объемов данных можно использовать как базы данных, так и библиотеки языков программирования (например Pandas). Ниже рассмотрим несколько вариантов обработки.

SQL — это наиболее распространенный язык манипулирования данными (DML), используемый для извлечения и управления данными в реляционной базе данных. Он был разработан для выполнения процедур базы данных (например, объединение, выбор, фильтрация, группировка) и мелкозернистого CRUD (создание, чтение, обновления и удаления). Таким образом, использование SQL — это необходимый первый шаг для преобразования исходных данных в более пригодный для использования формат.

База данных предоставляет те вещи, в которых фактически может не быть надобности, такие как параллелизм, блокировка, ограничения и прочие, которые требуют дополнительного времени для выполнения. В последствие может понадобиться заниматься дополнительной работой и решать разные проблемы, например, такие, как ошибки конфигурации. Но на более фундаментальном уровне SQL также не предназначен для преобразования данных определенным образом, и существует множество операций, которые выходят за рамки возможностей SQL. SQL не подходит для описания наборов строк, например для выдачи такой информации и выполнения такой работы, как: взвешенное скользящее среднее, манипуляции со строками, регулярные выражения и сводные таблицы.

Pandas — это просто пакет для языка программирования, при использовании него не возникает таких же сложностей, как при использовании языка для работы с базами данных, например, при использовании SQL. Он с самого начала поставляется с оптимизированной скоростью и готов к работе сразу после установки с помощью pip.

Ниже рассмотрим подробнее плюсы и минусы разных подходов для обработки больших объемов данных:

Использование языка для работы с базами данных SQL может быть удобно тем, что в нем достаточно простой синтаксис и отладка.

Базы данных SQL позволяют оптимизировать запросы, использовать транзакции (это группа последовательных операций с базой данных, которая представляет собой логическую единицу работы с данными), резервное копирование или восстановление данных и возможность хранить наборы данных, объем которых превышает объем оперативной памяти на устройстве, при помощи которого происходит работа с базой.

В сравнении с ним pandas ограничен размером оперативной памяти, поскольку не предоставляет такой возможности. SQL работает с базами данных, которые оснащены достаточным объемом памяти для обработки огромных объемов информации.

В отличие от классических методов работы с базами данных, язык программирования Python предоставляет широкие возможности для решения задач статистического анализа данных, а также для построения моделей машинного обучения и искусственного интеллекта.  С появлением библиотеки Pandas для Python очень упростилась и ускорилась работа с данными, потому что с ее помощью упростились и появились разные функции и процедуры, используемые для этого, например:

* процедура загрузки/выгрузки данных из внешних источников;
* процедура очистки и предварительной обработки данных;
* встроенная поддержка категориальных данных;
* удобный интерфейс для дальнейшей передачи обработанных данных в модели машинного обучения и другие модули Python, работающие с данными;
* встроенная визуализация.

Если количество возвращаемых данных остается неизменным или увеличивается (например, при выполнении таких действий, как добавление сложно вычисляемых столбцов, перекрестные соединения и прочие), задачу лучше решать с помощью библиотеки Pandas.

Решая сложные, комплексные задачи, Pandas можно и нужно использовать совместно с SQL.

### Выбор среды для парсинга

Существует множество инструментов и языков программирования для очистки контента в интернете, при этом именно Python обеспечивает плавный и простой процесс. Веб-скрейпинг может быть выполнен на Python с помощью таких библиотек, как Requests, BeautifulSoup, Scrapy и Selenium.

Beautiful-Soup, является отличным инструментом для написания парсеров из-за широкого функционала. С его помощью она может моментально скачать информацию с любой страницы. Фреймворк поможет получить данные из документов формата XML, HTML.

### Выбор среды обработки данных

В работе была выбрана pandas в связи с удобствами использования Python и не самым большим объемом данных, к которому будет применяться мой алгоритм.

### Расчет корреляции

Корреляция - статистический коэффициент, отражающий взаимосвязь случайных величин. В случае если одна из величин изменяется, это должно приводить к изменению другой величины.

Для того, чтобы вручную найти коэффициент корреляции необходимо:

* собрать данные;
* вычислите среднее арифметическое;
* вычислите стандартное отклонение;
* вычислить коэффициент корреляции по формуле, представленной ниже:

(1)

где – разница среднего значения произведения признаков и произведения средних значений признаков,

– [стандартное отклонение](http://www.mathprofi.ru/formula_dispersii_standartnoe_otklonenie_koefficient_variacii.html#sko) признака x,

– [стандартное отклонение](http://www.mathprofi.ru/formula_dispersii_standartnoe_otklonenie_koefficient_variacii.html#sko) признака y.

# 2 ПРАКТИЧЕСКАЯ ЧАСТЬ

## Исходные данные

В качестве исходных данных было получено:

* федеральный справочник с кодами регионов;
* файл с контекстными данными, включающий в себя столбцы:

1. где расположена образовательная организация;
2. наименование населенного пункта;
3. размер населенного пункта;
4. ОО является специальной (коррекционной);
5. количество обучающихся в ОО (в сумме по всем классам);

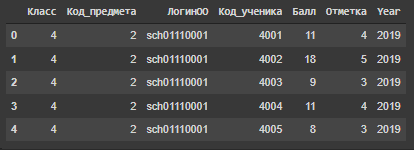
* csv файл в котором указаны:

1. класс;
2. код предмета;
3. логин ОО;
4. код ученика;
5. балл;
6. отметка;
7. год.

Рассмотрим исходные данные. База ВПР имеет поля:

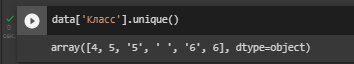
* класс, в котором учится учащийся;
* код предмета, который он сдает;
* логин образовательной организации, в которой он учится;
* балл, полученный по предмету;
* отметка, которая зависит от балла;
* год, в котором сдавался экзамен.

На рисунке 2.1 представлена база ВПР.



1. – База ВПР

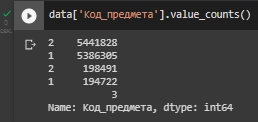
На рисунке 2.2 предоставлены данные по 4, 5 и 6 классам.



1. – Классы

Код предмета принимает значения от 1 и 2.

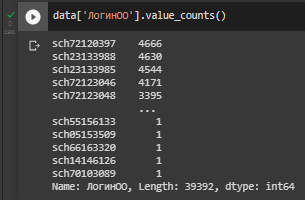
На рисунке 2.3 представлен код предмета.



1. – Код предмета

Логин ОО состоит из 11 символов, первые 2 цифры которого соответствуют номеру региона, в котором находится школа. В таблице предоставлены 39392 разных логинов и соответственно столько же школ.

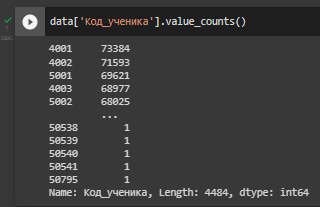
На рисунке 2.4 представлены логины.



1. – Логины

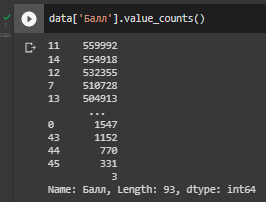
Код ученика представлен 4484 уникальными номерами.

На рисунке 2.5 представлен код ученика.



1. – Код

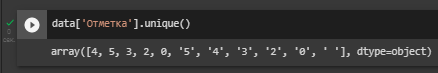
На рисунке 2.6 представлены данные по баллам.



1. – Балл

В зависимости от полученных баллов отметки разбиваются на 2, 3, 4, 5.

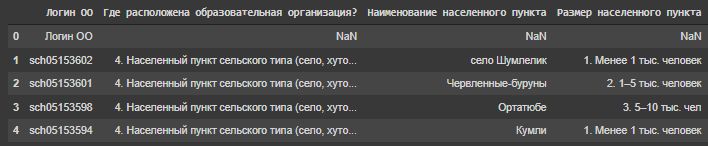
На рисунке 2.7 представлены данные по отметкам.



1. – Отметка

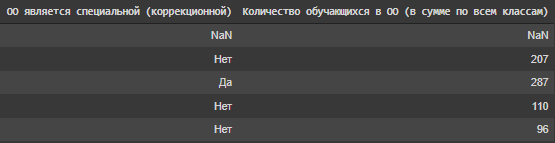
В базе с контекстными данными имеются данные, описывающие: Логин ОО; то, где расположена образовательная организация; наименование населенного пункта; размер населенного пункта; является ли специальной (коррекционной); количество обучающихся в ОО (в сумме по всем классам).

На рисунке 2.8 представлены контекстные данные.



1. – Контекстные данные 1

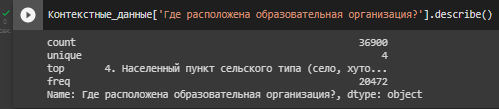
На рисунке 2.9 представлены контекстные данные.



1. – Контекстные данные 2

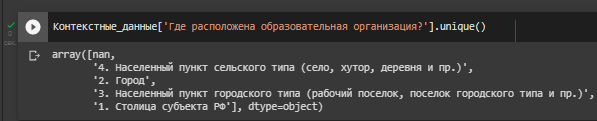
Колонка “Где расположена образовательная организация?” отображает тип местности, в которой находится ОО и включает четыре уникальных значения.

На рисунке 2.10 представлены типы местности.



1. – Типы местности 1

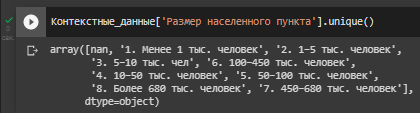
На рисунке 2.11 представлены типы местности.



1. – Типы местности 2

Размер населенного пункта включает в себя 8 позиций.

На рисунке 2.12 представлен размер населенного пункта.



1. – Размер населенного пункта

Для комплексной оценки необходимо большее количество контекстных данных, например: транспортная доступность, ВВП, количество культурных объектов, климатические условия, скорость интернета.

## Описание/представление результатов

### 2.2.1 Парсинг

Для сбора данных с веб-страницы необходимо:

* Проверить веб-страницу: нажав CTRL + U, можно просмотреть исходный код веб-страницы. Просматриваемый исходный код — это тот же самый код, из которого программа будет собирать необходимые данные. Обладая некоторыми базовыми знаниями в области HTML, можно проанализировать исходный код и найти HTML-разделы или элементы, которые содержат некоторые данные, такие как заголовки новостей, обзор новостей, дата статьи и так далее.
* Извлечь веб-страницу.

import requests

response = requests.get("адрес")

print(response.status\_code)

print(response.content)

При этом response.status\_code возвращает код ответа, указывающий, был ли запрос успешным или нет. Например, код состояния 200 указывает на то, что запрос был успешным, 4\*\* означает ошибку клиента, а 5\*\* означает ошибку сервера. Response.content возвращает содержимое ответа, которое является исходным кодом веб-страницы, и это тот же исходный код, который доступен вам при просмотре его в веб-браузере.

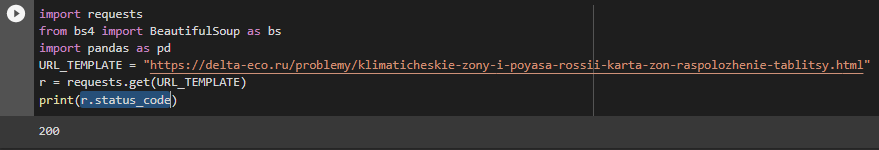
* Привести данные в необходимый вид.

Для парсинга был использован файл справочник, из него было взято соотношение регионов номерам и перенес в отдельный файл data.xls.

Была использована библиотеку Python BeautifulSoup. Она преобразует HTML и XML в синтаксические деревья, используя парсеры, такие как html5lib и Lxml, чтобы извлекать нужные данные.

Был использован метод get, который позволяет получить или извлечь некоторые данные из указанного ресурса. R.status\_code возвращает код, указывающий, был ли запрос успешным или нет. 200 означает успешный.

На рисунке 2.13 представлено использование библиотеки BeautifulSoup.



1. – BeautifulSoup

Доступ к исходному коду веб-страницы был получен с помощью метода .text:

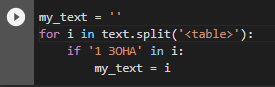
На рисунке 2.14 представлено использование поля text.



1. – Text

Split возвращает список слов в строке, где разделителем по умолчанию является пробел.

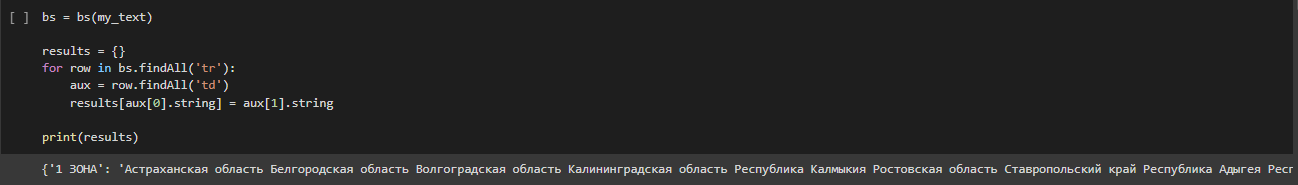
На рисунке 2.15 представлено использование метода split.



1. – Split

Метод .find\_all  BS просматривает и извлекает всех потомков тега, которые соответствуют переданным фильтрующим аргументам.

На рисунке 2.16 представлено использование метода find\_all.

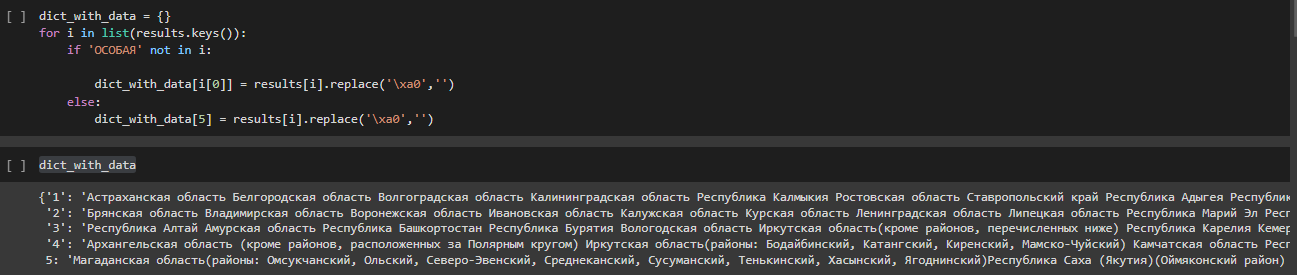


1. – Find\_all

На данный момент получены все регионы, распределенные по зонам в виде строки.

Приведу в более приличный вид.

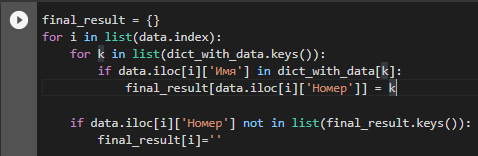
На рисунке 2.17 представлены данные Dict\_with\_data.



1. – Dict\_with\_data

Соотнесу номера регионов и номера температурных зон.

На рисунке 2.18 представлено Reg\_TimeZone.



1. – Reg\_TimeZone

В результате получены 2 столбца: в первом находится номер региона, во второй – номер температурной зоны, в которой он находится.

### Обработка данных

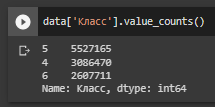
Начальные данные из исходной таблицы выглядят не очень хорошо.

На рисунке 2.19 представлены начальные данные.



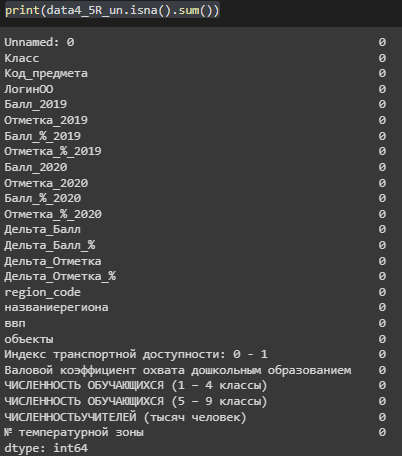
1. – Начальные данные 1

На рисунке 2.20 представлены начальные данные.



1. – Начальные данные 2

На рисунке 2.21 представлена проверка на наличие пустых значений.



1. – Проверка на наличие пустых значений

### Результаты расчета корреляции

Корреляции для результатов 4-х классов (математика без школ с динамикой роста больше 30%)

В таблицах 2.1 и 2.2 представлены корреляционные данные (округлены до третьего знака после запятой).

Таблица 2.1 – Корреляционные данные 1 часть

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Обеспеченность учителями | ВРП | Культурные объекты | Отметка\_2020 |
| Балл | -0.054 | 0.093 | 0.060 | 0.934 |

Таблица 2.2 – Корреляционные данные 2 часть

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ЧИСЛЕННОСТЬ ОБУЧАЮЩИХСЯ (1 – 4 классы) | ОБУЧАЮЩИЕСЯ (тыс чел) | ЧИСЛЕННОСТЬ УЧИТЕЛЕЙ (тысяч человек) | № температурной зоны | Валовой коэффициент охвата дошкольным образованием |
| Балл | 0.027 | 0.014 | 0.002 | 0.009 | 0.199 |

В таблицах 2.3 и 2.4 ниже представлены корреляционные данные по регионам.

Таблица 2.3 – Корреляционные данные по регионам 1 часть

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Обеспеченность учителями | ВРП | Культурные объекты | Индекс транспортной доступности |
| Отметка | -0.159 | 0.051 | 0.218 | 0.558 |

Таблица 2.4 – Корреляционные данные по регионам 2 часть

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ЧИСЛЕННОСТЬ ОБУЧАЮЩИХСЯ (1 – 4 классы) | ОБУЧАЮЩИЕСЯ (тыс чел) | ЧИСЛЕННОСТЬ УЧИТЕЛЕЙ (тысяч человек) | № температурной зоны | Валовой коэффициент охвата дошкольным образованием |
| Отметка | 0.024 | 0.074 | 0.051 | -0.082 | 0.517 |

# ЗАКЛЮЧЕНИЕ

В выпускной квалификационной работе бакалавра был проведен анализ корреляций результатов различных оценочных процедур в образовательных организациях с различными контекстными данными.

Была изучена документацию для работы с pandas и beautifulsoup4;

Были выделены контекстные данные, которые влияют на успеваемость;

Создать dataframe с идентификаторами школ и их контекстными данными, для дальнейшего использования в кластеризации школ;

Рекомендации по использованию и развитию разработанного подхода:

Работа, проделанная в ВКР, используется в дальнейшем для кластеризации школ с помощью алгоритмов иерархической, двухэтапной кластеризации, а также k-means и c-means; что служит основой для предсказаний успеваемости с помощью методов машинного обучения.

Для улучшения подхода можно использовать большее количество критериев для создания кластеров и увеличения точности оценки качества образования. Большим шагом в развитии было бы привязать критерии к геолокации самих школ, а не только к самим регионам. Также важная часть развития – автоматизированный сбор всех необходимых данных с использованием искусственного интеллекта.

Реализованные технологии получения данных для ОО планируется применить для получения таких переменных как: кол-во организаций, осуществляющих дополнительное образование, обеспеченность природными ресурсами; кол-во ВУЗов и СПО в регионе; для расширения и уточнения зависимости успеваемости ОО от контекстных факторов.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Ястребов Г.А., Пинская М.А., Косарецкий С.Г., Использование контекстных данных в системе оценки качества образования: опыт разработки и апробация инструментария // Вопросы образования. - 2014.
2. Чудинский Р.М., Быканов А.С., Володин А.А., Малев В.В., «Исследование зависимости предметных результатов обучающихся 5-9 классов от контекстных данных общеобразовательных организаций воронежской области» // [Международный научно-исследовательский журнал](https://cyberleninka.ru/journal/n/mezhdunarodnyy-nauchno-issledovatelskiy-zhurnal). - 2020.
3. Кравцов С.С., Музаев А. Ахмедович, Основные подходы к анализу результатов национальных исследований качества образования // [Педагогические измерения](https://cyberleninka.ru/journal/n/pedagogicheskie-izmereniya). - 2018.
4. Бастрикина В.В., «Проектирование веб-скрапера для получения данных с сайтов книжных издательств» // [Актуальные проблемы авиации и космонавтики](https://cyberleninka.ru/journal/n/aktualnye-problemy-aviatsii-i-kosmonavtiki). - 2018.
5. Документация Google Colab. [Электронный ресурс] – URL: <https://colab.research.google.com/?hl=in#scrollTo=-Rh3-Vt9Nev9>(дата обращения 12.04.2022)
6. Документация BeautifulSoup4. [Электронный ресурс] – URL: <https://www.crummy.com/software/BeautifulSoup/bs4/doc.ru/bs4ru.html>(дата обращения 16.04.2022)
7. ЕМИСС Государственная Статистика. [Электронный ресурс] – URL: <https://www.fedstat.ru/indicator/33379>(дата обращения 18.04.2022)
8. Федеральная служба государственной статистики. Регионы России. Социально-экономические показатели. [Электронный ресурс] – URL: <https://gks.ru/bgd/regl/b20_14p/Main.htm>(дата обращения 18.04.2022) – 2020.
9. Климатические зоны и пояса России. [Электронный ресурс] – URL:

<https://delta-eco.ru/problemy/klimaticheskie-zony-i-poyasa-rossii-karta-zon-raspolozhenie-tablitsy.html>(дата обращения 18.04.2022)

1. Транспортная доступность как индикатор развития региона. [Электронный ресурс] – URL: https://ecfor.ru/wp-content/uploads/2019/11/transportnaya-dostupnost-kak-indikator-razvitiya-regiona.pdf(дата обращения 18.04.2022)

# ПРИЛОЖЕНИЕ А

**Код программы**

Работа с dataframe

Подключение Pandas и Google Диск

import pandas as pd

from google.colab import drive

drive**.**mount('/content/drive')

Чтение csv

data **=** pd**.**read\_csv('/content/drive/MyDrive/pupils\_ruma456\_2019\_2020.csv')

Очистка данных

data **=** data[data['Отметка'] **!=** ' ']

data[['Балл']] **=** data[['Балл']]**.**apply(pd**.**to\_numeric)*##количественные*

data[['Класс', 'Код\_предмета', 'Код\_ученика', 'Отметка', 'Year']] **=** data[['Класс', 'Код\_предмета', 'Код\_ученика', 'Отметка',

data[['Класс', 'Код\_предмета', 'Код\_ученика', 'Отметка', 'Year']] = data[['Класс', 'Код\_предмета', 'Код\_ученика', 'Отметка', 'Year']].astype(str)#качественные

data**.**head()

data['Класс']**.**value\_counts()

data['Отметка']**.**value\_counts()

data['Отметка']**.**unique()

data[data['Отметка'] **==** '0']**.**head(10000)

data[data['Отметка'] **==** "3"]**.**describe()

'''count - кол-во записей по

mean - среднее значение атрибута (15)

std - корень дисперсии выборки (можно сравнить разброс выборки чем больше тем больше разброс чем меньше тем ближе к ср знач)

min - самое мал значение баллов

50%- медиана

25%

75% квартили

max - максимум

мы проверили что на данный момент нельзя сделать однозначное отображение балла в оценку

'''

data[(data['Отметка'] **==** "4") **&** (data['Код\_предмета'] **==** "2") ]**.**describe()

data[(data['Отметка'] **==** "3") **&** (data['Код\_предмета'] **==** "2") ]**.**describe()

data['Отметка']**.**value\_counts()

def get\_slice(x):

return x[3:5]

get\_slice([1,2,3,4,5,5])

Out[ ]:

[4, 5]

data['region\_code'] **=** data['ЛогинОО']**.**apply(get\_slice)

data**.**head(100000000000000000000000000000)

data[['region\_code']] **=** data[['region\_code']]**.**apply(pd**.**to\_numeric)

Контекстные\_данные **=** pd**.**read\_excel('/content/drive/MyDrive/Конт\_данныеОО.xlsx')

Контекстные\_данные**.**head()

Контекстные\_данные**.**rename(columns**=**{'Логин ОО':'login',

'Где расположена образовательная организация?':'location\_type',

'Наименование населенного пункта':'location',

'Размер населенного пункта':'pop\_size',

'ОО является специальной (коррекционной)':'special\_flg',

'Количество обучающихся в ОО (в сумме по всем классам)':'student\_cnt'},

inplace**=**True)

Контекстные\_данные **=** Контекстные\_данные[1:]

*#костыль*

Контекстные\_данные['special\_flg']**.**value\_counts()

НЕТ 36103

ДА 797

Name: special\_flg, dtype: int64

Контекстные\_данные['special\_flg'] **=** Контекстные\_данные['special\_flg']**.**str**.**upper()

Контекстные\_данные['special\_flg']**.**value\_counts()

НЕТ 36103

ДА 797

Name: special\_flg, dtype: int64

Контекстные\_данные[Контекстные\_данные['student\_cnt']**==**"Руэм"]**.**value\_counts()

ob = pd.read\_excel('/content/drive/MyDrive/OBSchAYa.xlsx')

ob[['регион']] **=** ob[['регион']]**.**apply(pd**.**to\_numeric)

ob**.**tail(30)

tr **=** pd**.**read\_excel('/content/drive/MyDrive/трд1.xlsx')

tr**.**head(100)

tr[['reg\_num ']] **=** tr[['reg\_num ']]**.**apply(pd**.**to\_numeric)

ob**.**rename(columns**=**{'регион':'region\_code',},

inplace**=**True)

tr**.**rename(columns**=**{'reg\_num ':'region\_code',},

inplace**=**True)

data\_un **=** pd**.**merge(ob, tr, on**=**'region\_code', how**=**'inner')

data\_un**.**head(1000)

data\_un **=** data\_un**.**drop(['Регион','Ранг', 'Индекс транспортной доступности: 0 - 1'], axis**=**1)

data\_un**.**to\_excel("output1.xlsx")

Парсер

import requests

from bs4 import BeautifulSoup as bs

import pandas as pd

URL\_TEMPLATE = "https://delta-eco.ru/problemy/klimaticheskie-zony-i-poyasa-rossii-karta-zon-raspolozhenie-tablitsy.html"

r = requests.get(URL\_TEMPLATE)

print(r.status\_code)

from google.colab import drive

drive.mount('/content/drive')

text = r.text my\_text = ''

for i in text.split('<table>'):

    if '1 ЗОНА' in i:

        my\_text = i

bs = bs(my\_text)

results = {}

for row in bs.findAll('tr'):

    aux = row.findAll('td')

    results[aux[0].string] = aux[1].string

print(results)

dict\_with\_data = {}

for i in list(results.keys()):

    if 'ОСОБАЯ' not in i:

        dict\_with\_data[i[0]] = results[i].replace('\xa0','')

    else:

        dict\_with\_data[5] = results[i].replace('\xa0','')

dict\_with\_data

data = pd.read\_excel('data.xls')

data.head()

final\_result = {}

for i in list(data.index):

    for k in list(dict\_with\_data.keys()):

        if data.iloc[i]['Имя'] in dict\_with\_data[k]:

            final\_result[data.iloc[i]['Номер']] = k

    if data.iloc[i]['Номер'] not in list(final\_result.keys()):

        final\_result[i]=''

final\_result

# ПРИЛОЖЕНИЕ Б

**Дополненные данные после сбора и парсинга**

Таблица Б.1 – Данные 1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Unnamed: 0 | Unnamed: 0.1 | Класс | Код\_предмета | ЛогинОО | Балл\_2019 | Отметка\_2019 | Балл\_%\_2019 | Отметка\_%\_2019 |
| 0 | 0 | 0 | 4 | 2 | sch01110001 | 11,83 | 3,97 | 59,13 | 79,42 |
| 1 | 1 | 1 | 4 | 2 | sch01110003 | 13,37 | 4,06 | 66,83 | 81,15 |
| 2 | 2 | 2 | 4 | 2 | sch01110004 | 12,9 | 4,22 | 64,49 | 84,41 |

Таблица Б.2 – Данные 2

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Отметка\_2020 | Балл\_%\_2020 | Отметка\_%\_2020 | Дельта\_Балл | Дельта\_Балл\_% | Дельта\_Отметка | Дельта\_Отметка\_% | region\_code | Название региона | ввп |
| 4,05 | 60,92 | 80,92 | 0,35 | 1,79 | 0,08 | 1,5 | 1 | Республика Адыгея (Адыгея) | 309184,1 |
| 3,57 | 50,43 | 71,43 | -3,28 | -16,4 | -0,49 | -9,72 | 1 | Республика Адыгея (Адыгея) | 309184,1 |
| 3,42 | 46,34 | 68,36 | -3,63 | -18,15 | -0,8 | -16,05 | 1 | Республика Адыгея (Адыгея) | 309184,1 |

Таблица Б.3 – Данные 3

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| объекты | Индекс транспортной доступности: 0 - 1 | Валовой коэффициент охвата дошкольным образованием | ЧИСЛЕННОСТЬ ОБУЧАЮЩИХСЯ (1 – 4 классы) | ЧИСЛЕННОСТЬ ОБУЧАЮЩИХСЯ (5 – 9 классы) | ЧИСЛЕННОСТЬУЧИТЕЛЕЙ (тысяч человек) | № температурной зоны | ОБУЧАЮЩИЕСЯ (тыс чел) | обеспеченность учителями |
| 5 | 0,28 | 67,7 | 22 | 20 | 3,6 | 1 | 54,5 | 0,066055 |
| 5 | 0,28 | 67,7 | 22 | 20 | 3,6 | 1 | 54,5 | 0,066055 |
| 5 | 0,28 | 67,7 | 22 | 20 | 3,6 | 1 | 54,5 | 0,066055 |