МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра материаловедения, технологии
и управления качеством

ПРИМЕНЕНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ ПРИ СБОРЕ И АНАЛИЗЕ ДАННЫХ ДЛЯ КОНТРОЛЯ КАЧЕСТВА НА ПРИМЕРЕ СПИСКОВ АБИТУРИЕНТОВ ВУЗОВ САРАТОВА

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента магистратуры 2 курса 2301 группы
направления 27.04.02 «Управление качеством»,
профиль «Менеджмент качества в инженерной и образовательной деятельности»
института физики

Мироедова Дениса Алексеевича

Научный руководитель, доцент, к.ф.-м.н.		А.В. Козловский
должность, уч. степень, уч. звание	подпись, дата	инициалы, фамилия
Зав. кафедрой, д.ф.-м.н., профессор		С.Б. Вениг
должность, уч. степень, уч. звание	подпись, дата	инициалы, фамилия

Саратов 2024

Введение. Для контроля качества приема абитуриентов, а также для агрегации информации об абитуриенте на сайтах университетов существует **проблема автоматизированного сбора данных**. Процесс автоматического сбора и анализа данных из различных источников с помощью специальных программных инструментов называется парсинг. Парсинг может быть полезен во многих сферах деятельности, в том числе в маркетинге, бизнесе, исследованиях, медицине, науке и т.д. [1-2].

Парсинг позволяет автоматизировать сбор данных и получить информацию, которая может быть использована для принятия решений, определения трендов, анализа конкурентов и многого другого [3].

Также парсинг позволяет получить данные, которые могут быть приведены к более удобному для анализа виду. Например, можно извлечь только нужную информацию из большого количества данных и представить ее в виде таблиц или графиков [4].

В целом, актуальность парсинга заключается в том, что он предоставляет широкие возможности для получения данных и их анализа, что может привести к улучшению бизнес-процессов, принятию более эффективных решений и улучшению управления организацией [5].

Таким образом, цель магистерской работы: изучение возможностей языков программирования для автоматизированного сбора информации об абитуриентах вузов Саратова, а также проведение исследовательского анализа этих данных.

Задачи магистерской работы:

1. изучить требования рособрнадзора к сайту образовательной организации, а также структуру сайтов вузов Саратова;
2. освоить основы языка программирования Python;
3. написать программу для сбора данных из раздела «Ранжированные списки поступающих» вузов Саратова;
4. провести предобработку и исследовательский анализ полученных данных;

5. сделать выводы об эффективности автоматизированного сбора данных сайтов образовательных организаций.

Выпускная квалификационная работа занимает 65 страниц, имеет 24 рисунков и 8 таблиц.

Обзор составлен по 31 информационным источникам.

Во введение рассматривается актуальность работы, устанавливается цель и выдвигаются задачи для достижения поставленной цели.

Основное содержание работы

Первый раздел представляет собой теоретическая часть. В ней рассматривается правила приема и структура сайтов Саратовских университетов, таких как: СГУ, СГТУ и СГАУ.

Описываются основы языка программирования Python [6]:

1. Переменные: переменные используются для хранения данных. В Python они объявляются простым присваиванием значения.

2. Типы данных: Python поддерживает различные типы данных, включая числа (целые числа, с плавающей запятой), строки, списки, кортежи и словари.

3. Операторы: Python поддерживает различные математические операторы, такие как +, -, *, /, а также операторы сравнения и логические операторы.

4. Условные операторы: в Python используются операторы if, elif и else для выполнения определенного блока кода в зависимости от условия.

5. Циклы: Python поддерживает циклы for и while для выполнения повторяющихся операций.

6. Функции: в Python функции объявляются с использованием ключевого слова def, их можно вызывать в различных частях программы для выполнения определенных задач.

7. Модули: Python имеет множество встроенных модулей, которые можно импортировать для выполнения различных функций, например, модуль math для математических операций.

8. Обработка исключений: для обработки ошибок в Python используется механизм исключений с помощью ключевых слов try, except и finally.

9. Списки и словари: списки представляют упорядоченный набор элементов, а словари – неупорядоченные коллекции ключей и значений.

10. Классы и объекты: Python поддерживает объектно–ориентированное программирование, где классы используются для создания объектов с определенными свойствами и методами.

Для анализа данных использовались следующие инструменты контроля качества:

1. Контрольный листок (рисунок 1) – это инструмент качества, используемый для систематической проверки выполнения определенных задач, процессов или стандартов. Обычно контрольный листок представляет собой список критериев или шагов, которые необходимо выполнить, и позволяет отслеживать выполнение работы и выявлять отклонения от установленных стандартов [7].

Наименование документа	Контрольный листок по видам дефектов	
Предприятие: XXX Цех:_____ Участок:_____	Изделие: _____ Операция:_____ Контролер:_____	Кол-во деталей _____
Типы дефектов	Данные контроля	ИТОГО
Деформации	//// //// //// //// //// //// //// //// //	47
Царапины	//// //// //// //// //// //// //// //	42
Трещины	//// //// //// //// ////	24
Раковины	//// //// //// //// //// //// //	38
Пятна	//// //// //// //// //// //// //// //// ///	53
Разрыв	//// //	7
Прочие	//// //// //	12
Всего		

Рисунок 1 – Контрольный листок

2. Диаграмма Парето (рисунок 2) – это график, который используется для идентификации основных источников проблем или причин, которые приводят к негативным результатам. Диаграмма Парето помогает определить,

какие проблемы следует решать в первую очередь, чтобы достичь наибольшего улучшения [8].

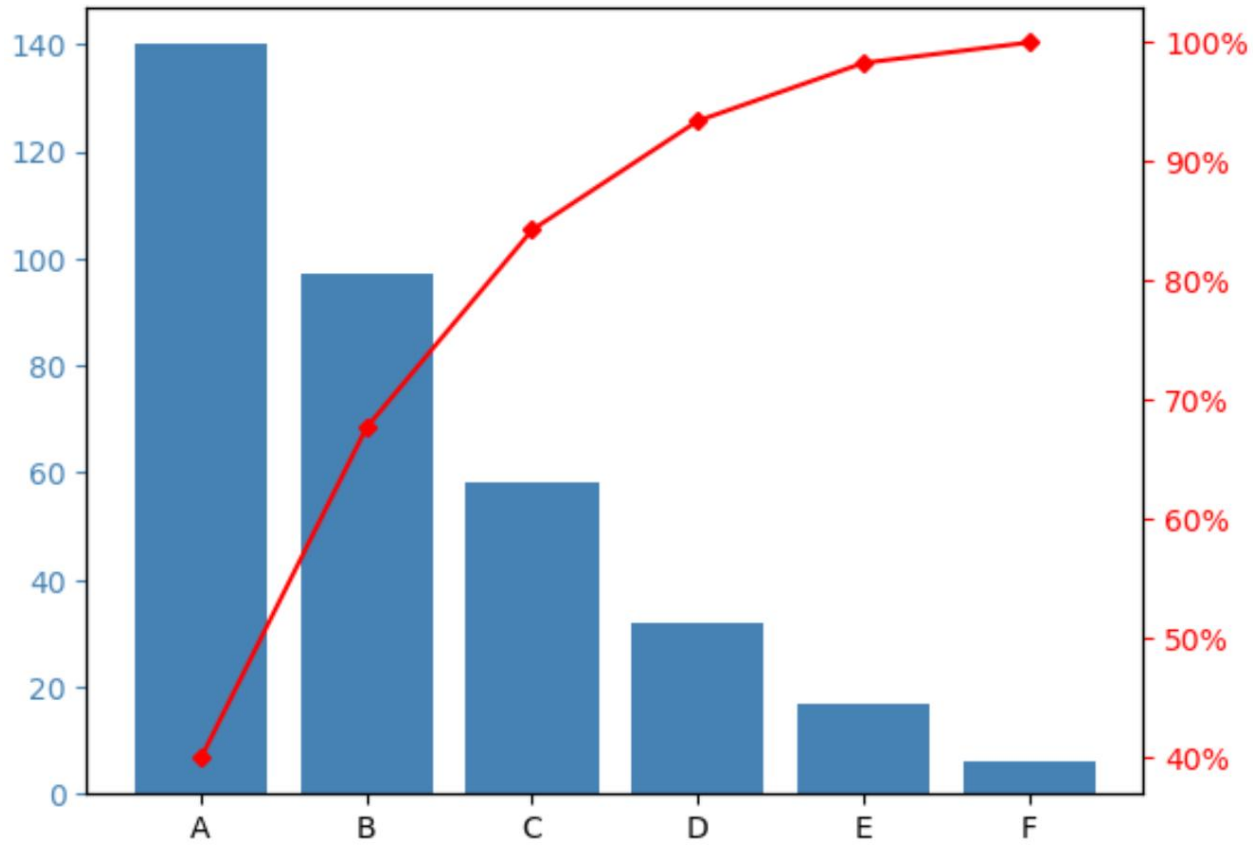

Рисунок 2 – Диаграмма Парето

3. Стратификация (расслоение) (рисунок 3) – это метод разделения данных на группы (страты) с целью более точного анализа. Стратификация помогает выявить различия между группами данных и принимать более обоснованные решения на основе этого анализа [9].

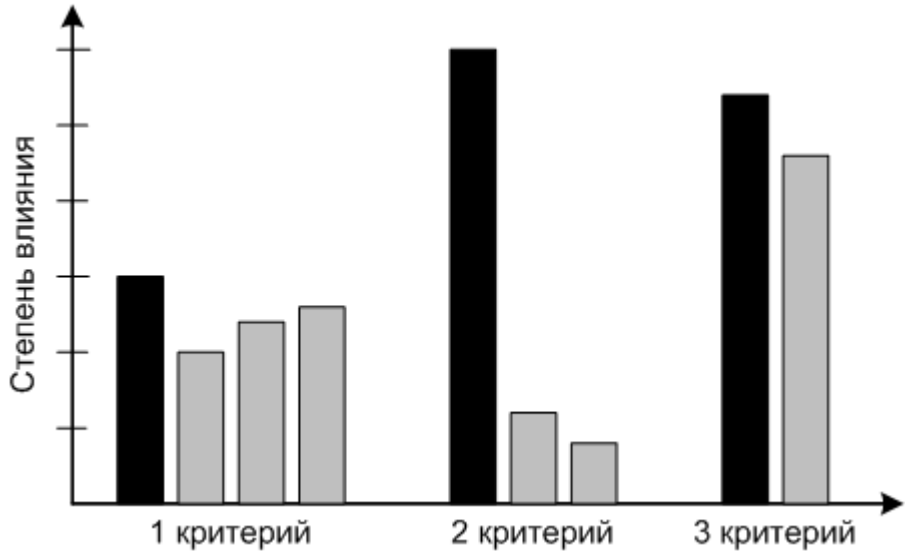

Рисунок 3 – Стратификация (расслоение)

4. Диаграмма Ишикавы (причинно-следственная диаграмма) (рисунок 4) – это графическое представление возможных причин проблемы или недостатка. Диаграмма Ишикавы включает различные категории потенциальных причин, такие как методы, материалы, оборудование, персонал и окружение, что помогает структурировать и анализировать информацию для выявления основных причин проблемы [10].

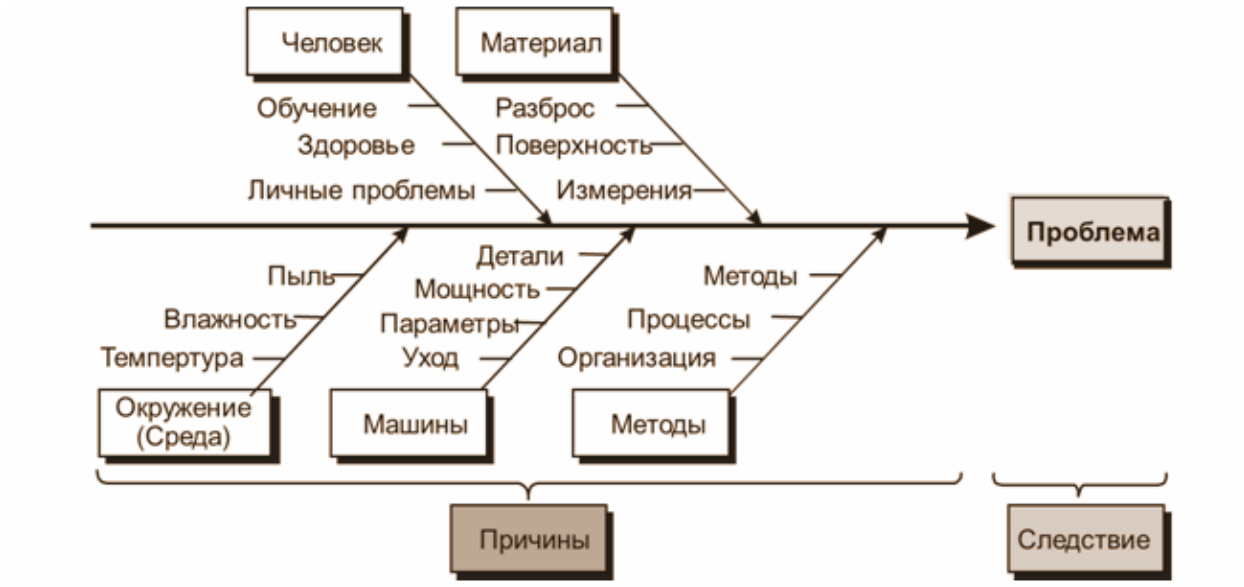

Рисунок 4 – Диаграмма Ишикавы (причинно-следственная диаграмма)

5. Контрольная карта (рисунок 5) – это инструмент мониторинга и управления качеством процесса. Контрольные карты используются для отслеживания переменных процесса, выявления отклонений от установленных

стандартов и принятия корректирующих мер для поддержания качества продукции или услуг на необходимом уровне [11].

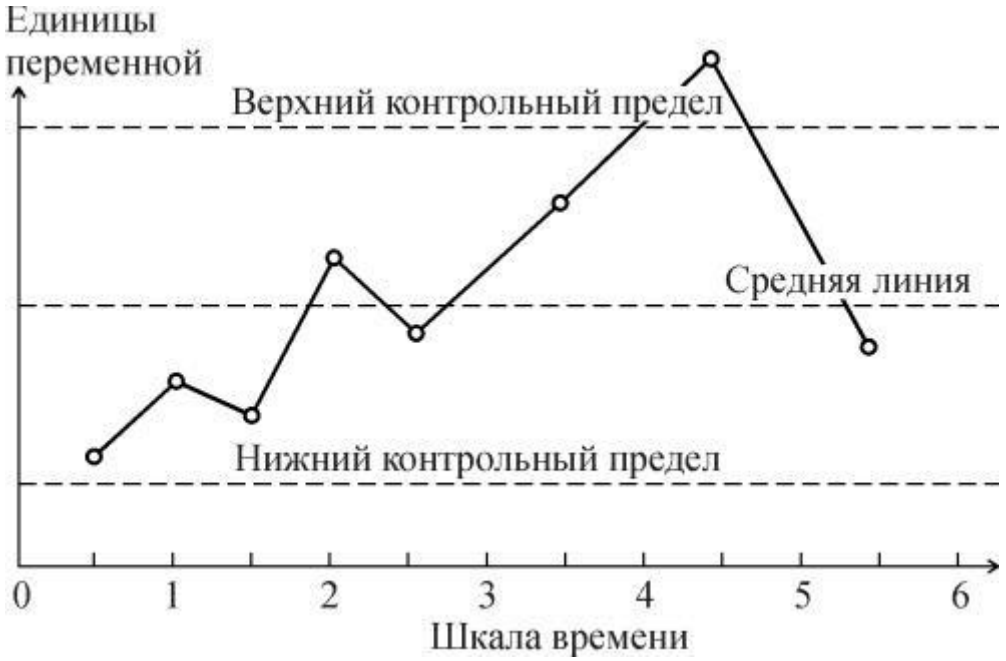

Рисунок 5 – Контрольная карта

6. Гистограмма (рисунок 6) это способ представления статистических данных в графическом виде − в виде столбчатой диаграммы. Она отображает распределение отдельных измерений параметров изделия или процесса. Иногда ее называют частотным распределением, так как гистограмма показывает частоту появления измеренных значений параметров объекта [12].

Высота каждого столбца указывает на частоту появлений значений параметров в выбранном диапазоне, а количество столбцов − на число выбранных диапазонов.

Важное преимущество гистограммы заключается в том, что она позволяет наглядно представить тенденции изменения измеряемых параметров качества объекта и зрительно оценить закон их распределения. Кроме того, гистограмма дает возможность быстро определить центр, разброс и форму распределения случайно величины. Строится гистограмма, как правило, для интервального изменения значений измеряемого параметра.

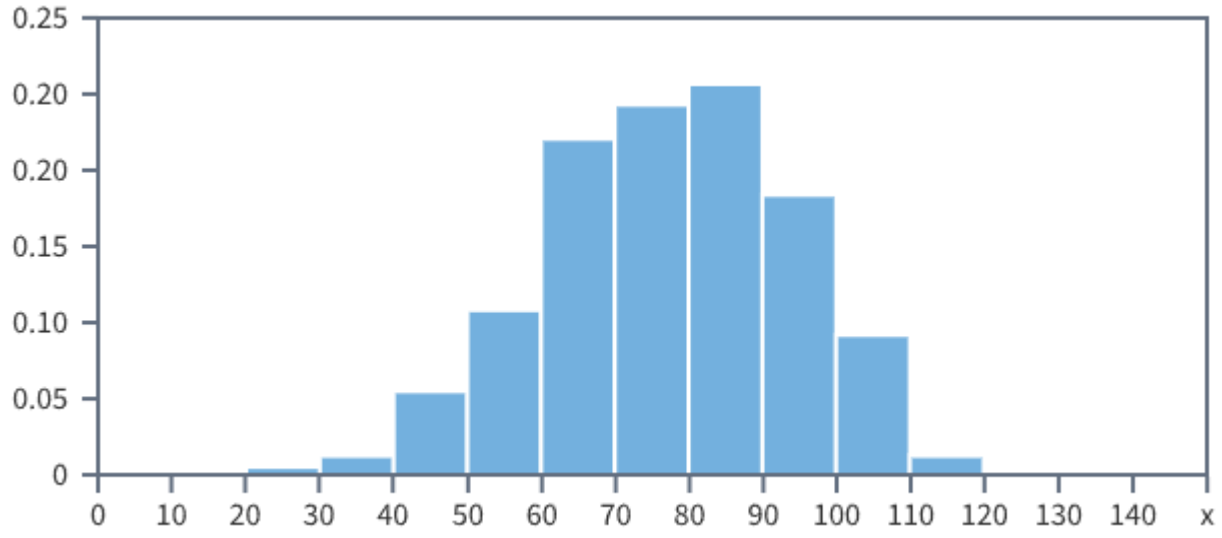

Рисунок 6 – Гистограмма

7. Диаграмма разброса (рисунок 7) – это график, который используется для иллюстрации отношения между двумя переменными. Он позволяет определить наличие корреляции или зависимости между этими переменными. Диаграмма разброса помогает выявить закономерности и взаимосвязи между данными [13].

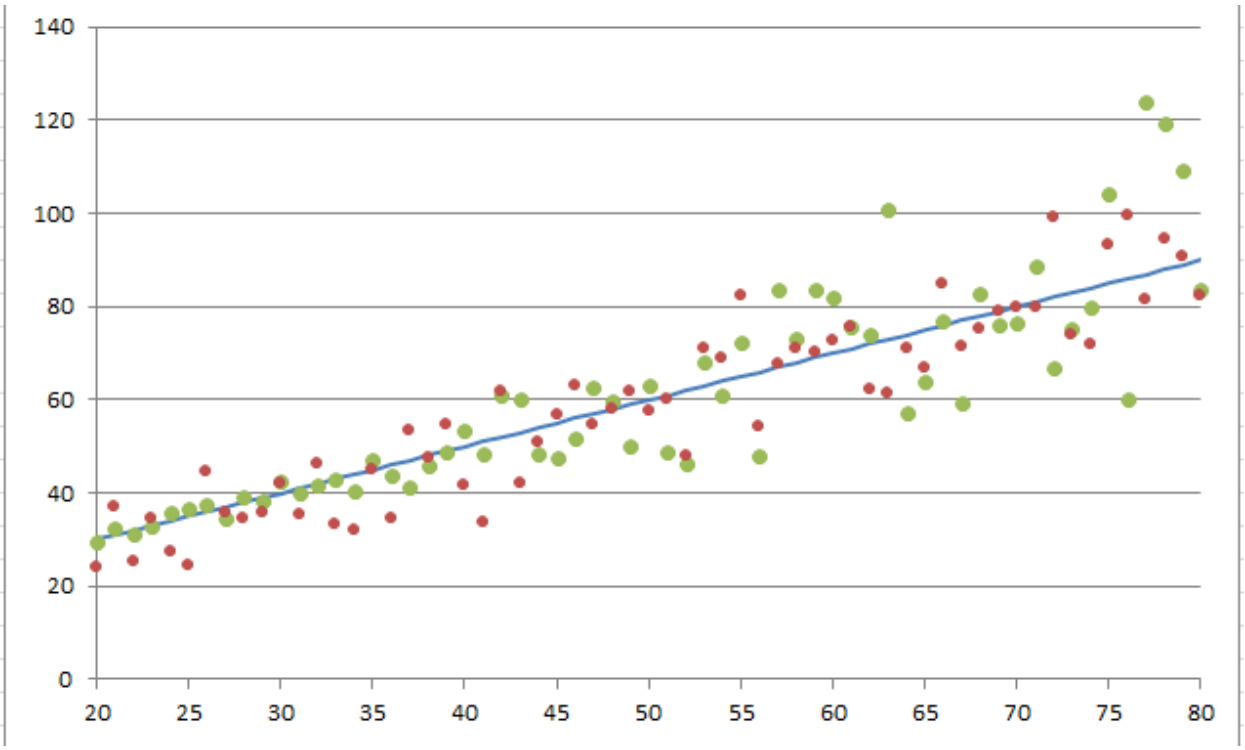

Рисунок 7 – Диаграмма разброса

Во втором разделе работы описывается практическая часть. В данном разделе производится сбор данных из раздела «Ранжированные списки поступающих» веб-сайтов образовательных учреждений. Пишется код для парсинга данных сайтов и предоставляется результат в качестве таблиц, где указано количество поданных заявлений за июль 2023 года. Следующим шагом будет предобработка и исследовательский анализ данных абитуриентов Саратова за 2023 год. На данной стадии строится гистограмма и такие диаграммы как: столбчатая диаграмма, диаграмма размаха.

1. Диаграмма размаха, или «ящик с усами» [14] – это графическое представление статистических данных, которое используется для визуализации распределения набора значений и выявления выбросов.

Квартили [15] – это значения, которые делят распределение на 4 равные части одинакового размера. Различают первый квартиль (Q1), второй (Q2 – это медиана) и третий (Q3). Первый квартиль – это такое значение, меньше которого будет 25% наблюдений, а 75% будут больше. Q2 является медианой и делит распределение пополам. Q3 (третий квартиль) – это значение, больше которого будет 25% наблюдений. Диаграмма размаха позволяет быстро оценить центральную тенденцию, разброс данных, наличие выбросов и симметрию распределения.

Строится диаграмма размаха следующим образом: на графике изображается прямоугольник (ящик), который показывает интерквартильный размах – разницу между верхним и нижним квартилями данных. Границами ящика служат квартили, а внутренняя линия – медиана. Верхняя и нижняя границы «усов» – это обычно 1,5 межквартильных размаха (длина ящика) от верхнего и нижнего квартилей соответственно. Все значения, выходящие за пределы «усов», отображаются индивидуальными точками или могут быть классифицированы как выбросы.

Диаграмма размаха очень полезна для сравнения распределения данных между разными группами или наборами значений, обнаружения необычных или экстремальных значений, а также выявления общего вида распределения

данных. Этот инструмент позволяет наглядно представить ключевые статистические показатели и оценить характеристики данных без необходимости изучения подробных числовых значений.

На рисунке 8 представлен пример диаграммы размаха.

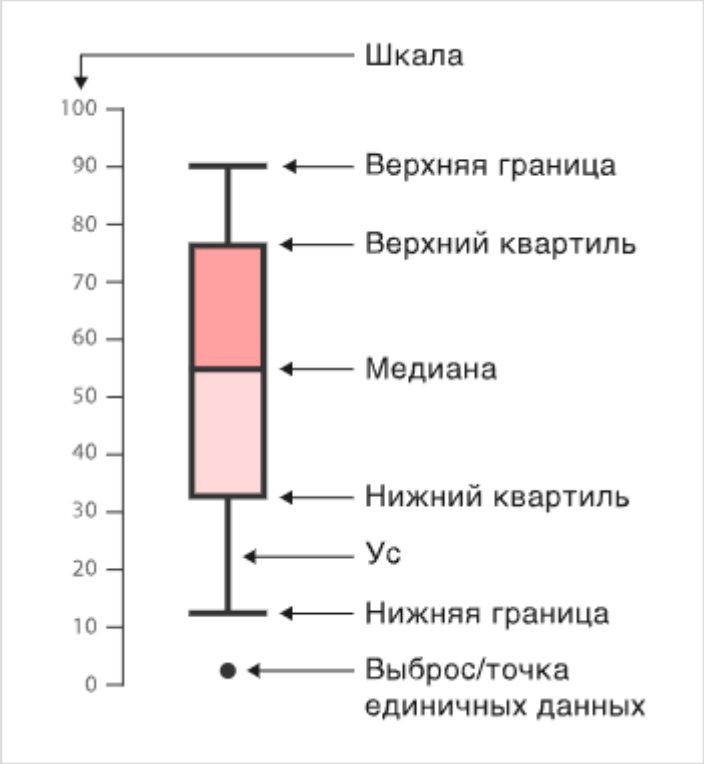

Рисунок 8 – Диаграмма размаха, или «ящик с усами»

Заключение. В ходе выполнения работы, были построены таблицы суммы поданных документов и с помощью них выведены столбчатые диаграммы, благодаря которым, удалось выявить направления и факультеты с наибольшим и наименьшим количеством поданных документов во время приемной комиссии города Саратов за 2023 год, а также гистограммы по сумме баллов для бакалавров.

Были построены диаграммы размаха по сумме баллов и среднему значению балла для бакалавров в течение приемной комиссии СГУ за 2023 год. С помощью неё были выявлены факультеты с наибольшими и наименьшими баллами, выбросы или отклонения, которые превышают 1 и 3 квартиль диаграммы.

На основе полученных данных удалось выявить несколько преимуществ парсинга. Во-первых, парсинг позволяет собирать большие объемы

информации за короткий период времени, что делает его эффективным инструментом для анализа сайтов образовательных учреждений. Во-вторых, парсинг позволяет автоматизировать процесс сбора данных, что увеличивает его точность и уменьшает вероятность ошибок. В-третьих, парсинг позволяет проводить анализ больших объемов информации и выявлять полезные тенденции и зависимости.

Список использованных источников

1 Парсинг данных с сайтов: что это и зачем он нужен [Электронный ресурс] // Ringostat [Электронный ресурс] : [сайт]. – URL: https://blog.ringostat.com/ru/parsing-dannyh-s-saytov-chto-eto-i-zachem-on-nuzhen/ (дата обращения: 05.05.2024). – Загл. с экрана. – Яз. рус.

2 Парсинг – что это? Виды и примеры парсинга [Электронный ресурс] // Spark.ru [Электронный ресурс] : [сайт]. – URL: https://spark.ru/startup/leadozvon/blog/74676/parsing-chto-eto-vidi-i-primeri-parsinga (дата обращения: 05.05.2024). – Загл. с экрана. – Яз. рус.

3 Парсинг: что это такое и как работает [Электронный ресурс] // Содействие занятости [Электронный ресурс] : [сайт]. – URL: https://www.tgu-dpo.ru/news/2023/06/07/chto-takoe-parsing-i-kak-on-rabotat/ (дата обращения: 05.05.2024). – Загл. с экрана. – Яз. рус.

4 Основы парсинга на Python: от Requests до Selenium [Электронный ресурс] // Хабр [Электронный ресурс] : [сайт]. – URL: https://habr.com/ru/companies/selectel/articles/754674/ (дата обращения: 30.05.2024). – Загл. с экрана. – Яз. рус.

5 Что такое парсинг и как правильно парсить [Электронный ресурс] // Calltouch. Blog [Электронный ресурс] : [сайт]. – URL: https://www.calltouch.ru/blog/chto-takoe-parsing/ (дата обращения: 05.06.2024). – Загл. с экрана. – Яз. рус.

6 Основы программирования на языке python [Электронный ресурс] // УРФУ [Электронный ресурс] : [сайт]. – URL:

https://elar.urfu.ru/bitstream/10995/28769/1/978-5-7996-1198-9_2014.pdf (дата обращения: 07.06.2024). – Загл. с экрана. – Яз. рус.

7 Контрольный листок [Электронный ресурс] // Цент дистанционного обучения и повышения квалификации [Электронный ресурс] : [сайт]. – URL: https://de.donstu.ru/CDOCourses/structure/Prib_i_Tech_Reg/uprav_ka/904/4_1.html#:~:text=Обычно%20контрольный%20листок%20представляет%20собой%2Cданные%20без%20их%20последующего%20переписывания (дата обращения: 05.06.2024). – Загл. с экрана. – Яз. рус.

8 Диаграмма Парето [Электронный ресурс] // GanttPro [Электронный ресурс] : [сайт]. – URL: https://blog.ganttpro.com/ru/diagramma-pareto-chart-i-effektivnoe-upravlenie-proektami/ (дата обращения: 05.06.2024). – Загл. с экрана. – Яз. рус.

9 Стратификация [Электронный ресурс] // Инновации и бизнес [Электронный ресурс] : [сайт]. – URL: https://inbsn.ru/business-optimization/stratification.html#:~:text=Стратификация%20 (дата обращения: 05.06.2024). – Загл. с экрана. – Яз. рус.

10 Диаграмма исикавы [Электронный ресурс] // Белгородский машиностроительный техникум [Электронный ресурс] : [сайт]. – URL: https://bmt31.ru/wp-content/uploads/2021/03/DIAGRAMMA-ISIKAVY.pdf (дата обращения: 05.06.2024). – Загл. с экрана. – Яз. рус.

11 Контрольная карта Шухарта [Электронный ресурс] : свободная энциклопедия / текст доступен по лицензии Creative Commons Attribution-ShareAlike ; Wikimedia Foundation, Inc, некоммерческой организации. − Электрон. дан. (1984743 статей, 7966275 страниц, 252098 загруженных файлов). − Wikipedia®, 2001-2024. − URL: https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BD%D1%82%D1%80%D0%BE%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F_%D0%BA%D0%B0%D1%80%D1%82%D0%B0_%D0%A8%D1%83%D1%85%D0%B0%D1%80%D1%82%D0%B0 (дата обращения: 05.06.2024). − Загл. с экрана. − Последнее изменение страницы: 17:57, 4 февраля 2023 года. − Яз. рус.

12 Гистограмма [Электронный ресурс] // Менеджмент качества [Электронный ресурс] : [сайт]. – URL: https://www.kpms.ru/Implement/Qms_Histogram.htm (дата обращения: 08.06.2024). – Загл. с экрана. – Яз. рус.

13 Диаграмма разброса [Электронный ресурс] // Менеджмент качества [Электронный ресурс] : [сайт]. – URL: https://www.kpms.ru/Implement/Qms_Scatter_Diagram.htm (дата обращения: 08.06.2024). – Загл. с экрана. – Яз. рус.

14 Диаграмма размаха («ящик с усами») [Электронный ресурс] // Каталог визуализации данных [Электронный ресурс] : [сайт]. – URL: https://datavizcatalogue.com/RU/metody/diagramma_razmaha.html (дата обращения: 08.06.2024). – Загл. с экрана. – Яз. рус.

15 Основы статистики [Электронный ресурс] // Анализ данных, статистика и маркетинговые исследования [Электронный ресурс] : [сайт]. – URL: https://www.tidydata.ru/rebelytics/basic-statistics (дата обращения: 08.06.2024). – Загл. с экрана. – Яз. рус.