Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«Уральский федеральный университет имени первого Президента России Б.Н. Ельцина»

Физико-технологический институт

Кафедра теоретической физики и прикладной математики

ПРОЕКТ ПО МОДУЛЮ

по теме: Анализ данных

Руководитель:Шориков А.О.

Студент:Здерев П.А.

Группа:Фт-210005

Екатеринбург

2023

Содержание

1 Введение

В настоящее время из-за постоянного развития технологий становится всё более актуальной проблема анализа данных. Люди генерируют большое количество данных ежедневно – любимые песни, предпочтения в кино, цели в жизни. Данные больше не ограничиваются технологическими компаниями, сейчас они используются в абсолютно разных сферах для построения маркетинговых стратегий, улучшения качества обслуживания, сбора информации о пользователях данных. Согласно данным сайта IBM на 2019 год каждый день создается 2,5 квинтиллиона байт данных, а анализируется только 0,5% генерируемых данных!

Актуальность данной работы состоит в том, что в нашем быстро развивающемся мире важно, чтобы мы анализировали весь этот поток данных правильно и принимали соответствующие решения, однако хороших специалистов в данной сфере крайне мало, поэтому нам важно самим научиться разбираться в данных.

«Анализ данных» - это область информатики, занимающаяся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных данных. «Анализ данных» можно назвать искусством по причине того, что аналитик сам выбирает путь к получению выгоды из данных, исходя из своих интересов.

Объектом исследования выступают основные способы анализа данных.

Предметом исследования является две базы данных разного размера.

Целью работы является научиться базово проводить анализ на разных размерах базы данных.

Реализация данной цели исследования обусловила необходимость решения следующих **задач**:

– рассмотреть понятие, принципы и особенности анализа данных

– изучить методы и задачи анализа данных

– проанализировать корреляцию параметров в базах данных

– определить ошибки, которые можно допустить при некачественном анализе данных

– сделать тезисные выводы и предложения по результатам проведенного исследования.

Метод исследования: изучение и анализ двух различных баз данных.

2 Основная часть

2.1 В чем заключается анализ данных?

После успешного сбора точных и надежных данных, следующим шагом является извлечение соответствующей и полезной информации, скрытой в данных, для дальнейшего манипулирования и интерпретации. Процесс выполнения определенных расчетов и оценки с целью извлечения необходимой информации из данных называется анализом данных. Анализ данных может состоять из нескольких этапов, чтобы прийти к определенным выводам. Простые данные могут быть организованы очень легко, в то время как сложные данные требуют соответствующей обработки. Слово "обработка" означает переработку и обработку данных с целью их подготовки к анализу. Слово "анализ" относится к тесно связанной операции, которая выполняется с целью обобщения собранных данных и организации их таким образом, чтобы получить ответ на поставленные вопросы. Проще говоря, это означает изучение данных для определения присущих им фактов. Термин анализ относится к такому процессу, который облегчает данные для операций, предназначенных для того, чтобы сделать выводы для дальнейших манипуляций. Анализ данных предполагает организацию данных в надлежащем виде. Проблема анализа данных варьируется от исследования к исследованию. В "комплексных операциях", которые называются анализом данных, участвуют семь этапов. Эти шаги следующие[1]:

1. Классификация и табулирование

2. Графическое представление

3. Мера местоположения

4. Измерение изменчивости

5. Измерение взаимосвязи

6. Оценка неизвестного

7. Проверка гипотезы

Классификация данных - это процесс упорядочивания в классы в соответствии с некоторым сходством или общими характеристиками. Классификация также называется категоризацией данных. Классификация может быть выполнена на основе качества или атрибута, такого как пол, цвет кожи, грамотность, красота, IQ. Этот тип классификации называется качественной классификацией.

Второй тип классификации - это качественная классификация, которая проводится на основе переменных, таких как рост, вес.

Третий тип классификации - географическая классификация, такая как деревня, район, город, городская и сельская местность.

Четвертый тип классификации называется хронологической классификацией, которая проводится на основе времени, как еженедельная, ежемесячная, ежегодная. После проведения классификации частота закрытий может быть преобразована в пропорцию или процент.

$$p(\text{процент}) = \frac{\text{Часть}}{\text{Целое}}.$$

Табулирование или частотное распределение. Техника представления количественных данных, таких как рост, вес, АД, температура и другие биологические характеристики, измеряемые на физических весах, в строках и колонках, называется табулированием. Табулирование также известно как частотное распределение

переменных. Основная цель табулирования - сжать данные и облегчить их сравнение. В табличной форме данных легко получить необходимую интерпретацию. Данные выглядят очень наглядно. Подготовка таблицы - это искусство, требующее квалифицированной работы с данными. Подготовка таблицы зависит от объема и характера данных.

2.1.1 Представление данных

Данные можно отобразить с помощью графика и диаграммы вместо классификации табуляции. Существует множество причин для построения графика. Самая привлекательная причина заключается в том, что один простой график говорит больше, чем двадцать страниц прозы. Графики представляют собой краткое изложение данных. Обычно предлагается, чтобы графическое представление рассматривалось перед формальным статистическим расчетом. Графики дают визуальное представление данных. Графики полезны при подгонке данных под модель. Рассматривая графики, можно легко понять данные. Распространенные типы графиков[2]:

1. Простая гистограмма
2. Множественная гистограмма
3. Круговая диаграмма
4. Гистограмма
5. Диаграмма рассеяния

1. *Простая столбиковая диаграмма* состоит из вертикальных или горизонтальных полос одинаковой ширины и длины, пропорциональной величине значения, которое они представляют.

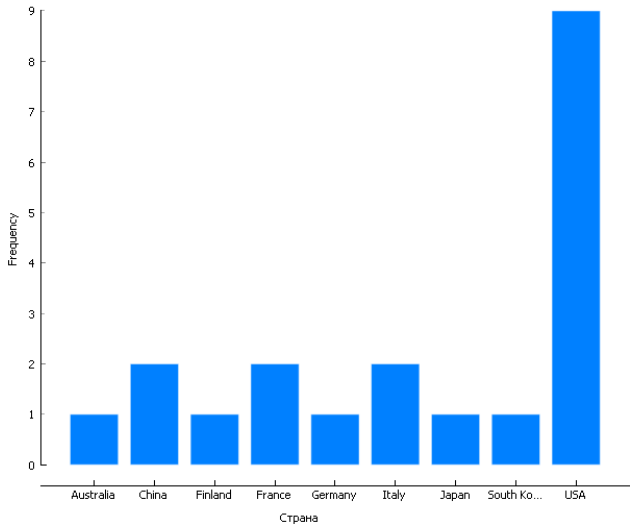

Рисунок 1: Простая столбиковая диаграммма

2. *Множественная или подразделенная диаграмма.* Это просто расширение простой гистограммы, которая представляет более чем один связанный набор данных.

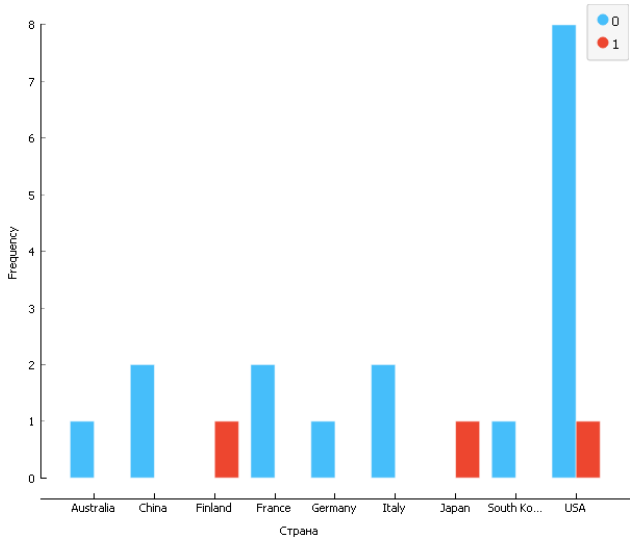

Рисунок 2: Множественная диаграмма.

3. *Круговая диаграмма* состоит из круга, который разделен на сектора, площадь которых пропорциональна различным компонентам от общего количества.

$$p(\text{частичный угол}) = \frac{\text{Часть}}{\text{Целое}} \cdot 360°.$$

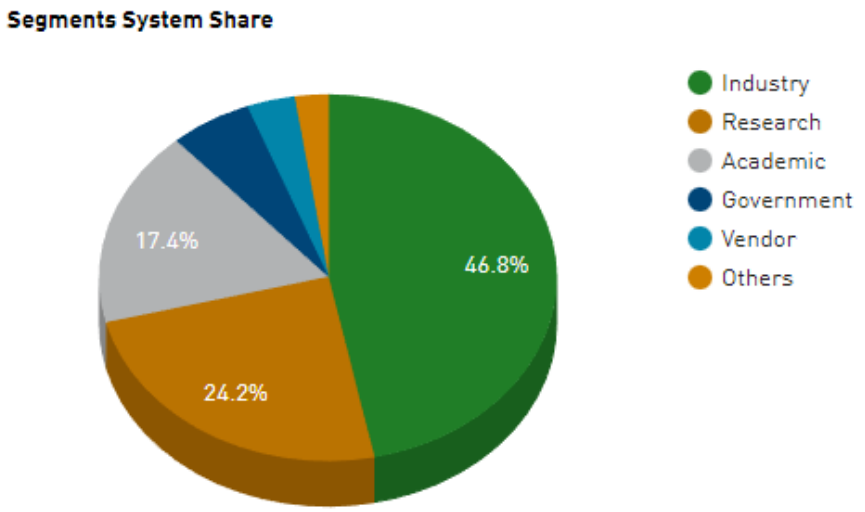

Рисунок 3: Круговая диаграмма.

4. *Гистограмма* - это график непрерывных данных, таких как вес, рост, возраст и т.д., позволяющий увидеть теоретическую форму данных. Кривая гистограммы говорит нам, являются ли данные перекошенными или симметричными. Гистограмма состоит из ряда смежных прямоугольников, построенных для сгруппированной частоты.

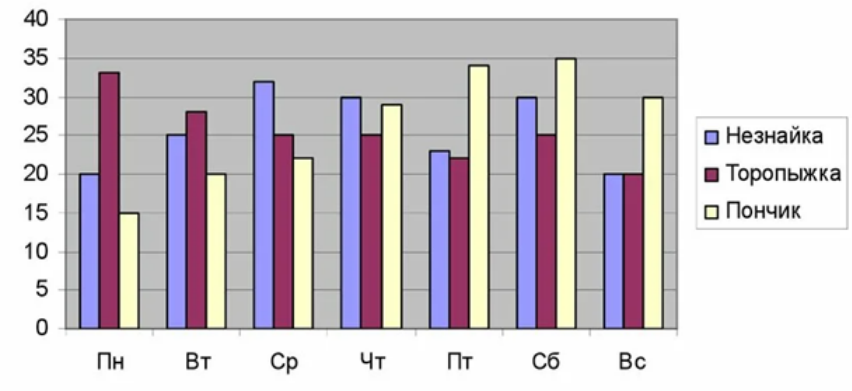

Рисунок 4: Гистограмма.

5. *Диаграмма рассеяния* - это набор пунктирных точек для представления отдельных фрагментов данных по горизонтальной и вертикальной оси.

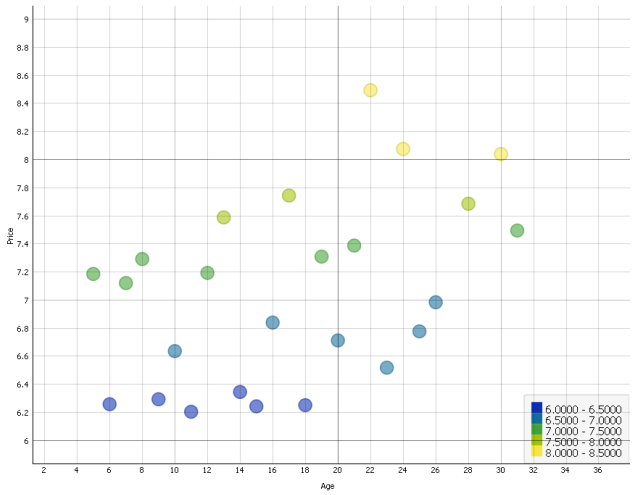

Рисунок 5: Диаграмма рассеяния.

2.1.2 Связь между данными

В определенных ситуациях исследователю интересно выяснить взаимосвязь между переменными. Существует ли сильная связь между двумя переменными или более слабая? Мера взаимосвязи классифицируется как: 1. Регрессия. 2. Корреляция. Регрессионный анализ используется для прогнозирования или оценки одной переменной на основе другой переменной. В регрессии мы намерены описать зависимость одной величины от другой. Линия регрессии математически определяется следующим образом: $y = a + bx$, где b - наклон линии, который измеряет изменение зависимой переменной при изменении независимой переменной на единицу, а «a» - начальное значение зависимой переменной.

Корреляция же описывает взаимосвязь между двумя переменными. Коэффициент корреляци измеряет степень взаимосвязи между двумя переменными и математически определяется как:

$$r = \frac{\sum xy - n\overline{xy}}{\sqrt{(\sum x^2 - nx^{-2})(\sum y^2 - ny^{-2})}}$$

.

Значение r лежит между -1 и +1.

2.1.3 Задача линейной регрессии

Теперь расскажем об самом базовом и основном методе анализе данных *линейной регрессии.* Линейная регрессия - это алгоритм, который обеспечивает линейную зависимость между независимой переменной и зависимой переменной для прогнозирования исхода будущих событий. Это статистический метод, используемый в науке о данных и машинном обучении для прогностического анализа[3].

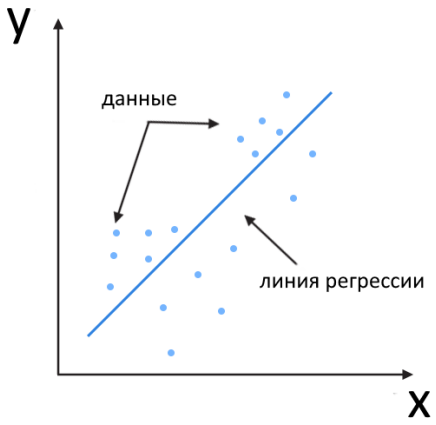

Рисунок 6: Модель линейной регрессии.

На приведенном выше рисунке,

Ось y = независимая переменная

Ось x = выходная / зависимая переменная

Линия регрессии = линия наилучшего соответствия модели

Здесь для заданных точек данных строится линия, которая соответствующим образом соответствует всем данным. Следовательно, это называется «линией наилучшего соответствия». Цель алгоритма линейной регрессии - найти эту линию наилучшего соответствия, показанную на рисунке выше.

Решением задачи линейной регрессии является найденное уравнение линейной регрессии, задающей линию наилучшего соответсвия. Данное уравнения для нескольких параметров принимает вид:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n,$$

где $\beta_0, \beta_1, \beta_2, ..., \beta_n$ - коэффициенты параметров регрессии.

Коэффициенты линейной регрессии показывают скорость изменения зависимой переменной по данному параметру, при фиксированных остальных параметрах, т.е. :

$$b_i = \frac{\partial y}{\partial x_i}, \qquad i = \overline{1, n}$$

.

Таким образом, при решенной задачи регрессии, мы можем предсказывать значение y, просто подставляя значения параметров x. Рассмотрим решение данной задачи в следующих примерах.

3 Анализ двух баз данных

3.1 Анализ большой базы данных

3.1.1 Предисловие

Для примера базового анализа данных выполним анализ рейтинга самых лучших суперкомпьютеров - сайт *top500.org*. Данный сайт имеет большое влияние в суперкомпьютерной индустрии и после каждого выхода новой редакции рейтинга многие выполняют различные подсчеты и публикуют суждения, основанные на результатах таких подсчетов. Довольно часто подсчеты посвящены вычислению различных долей в списке Top500 — например, вычисляют, какие доли приходятся на различные области применения суперкомпьютеров из Top500, или какие доли приходятся на суперкомпьютеры, использующие те или иные микропроцессоры. Анализируют и другие процентные распределения: доли различных архитектур, доли производителей суперкомпьютеров, доли стран и т.п. Анализ делают и сами создатели рейтинга, делая приложение к публикации рейтинга в виде постера с основной статистикой[4][5].

Удобный доступ к данным рейтинга доступен на сайте в виде *Excel*-таблицы. В таблице 500 строчек и 38 столбцов, получается 19 000 записей. Понятно, что вручную выполнить тонкий анализ такого количества данных невозможно. Поэтому автор будет использовать программу **Orange**, предназначенную для анализа данных. Все иллюстрации и все данные для расчетов в данной работе подготовлены при помощи этой программы.

3.1.2 Статистика

Проанализируем, какие страны лидируют в суперкомпьютерной индустрии, для этого построим диаграмму:

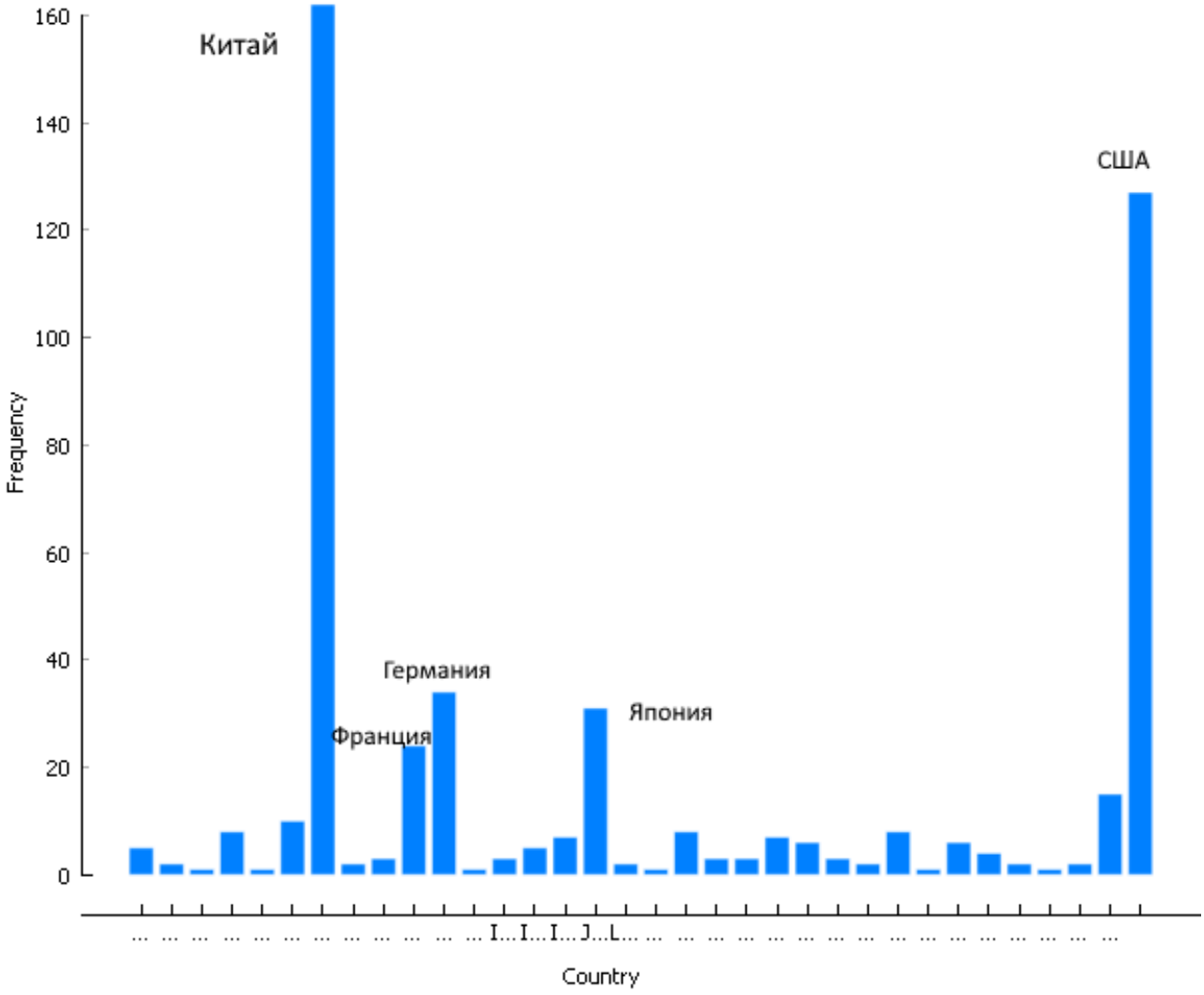

Рисунок 7: Диаграмма распределения стран в top500

Если поторопиться, то можно сделать вывод, что *в суперкомпьютерной индустрии есть пятерка лидеров: Китай, Франция, Германия, Япония, США,*

среди которых лидирует Китай. Однако, если сделать множественную гистограмму по тому, входит ли суперкомпьютер данной страны в 50 лучших, то мы получим следующую диаграмму:

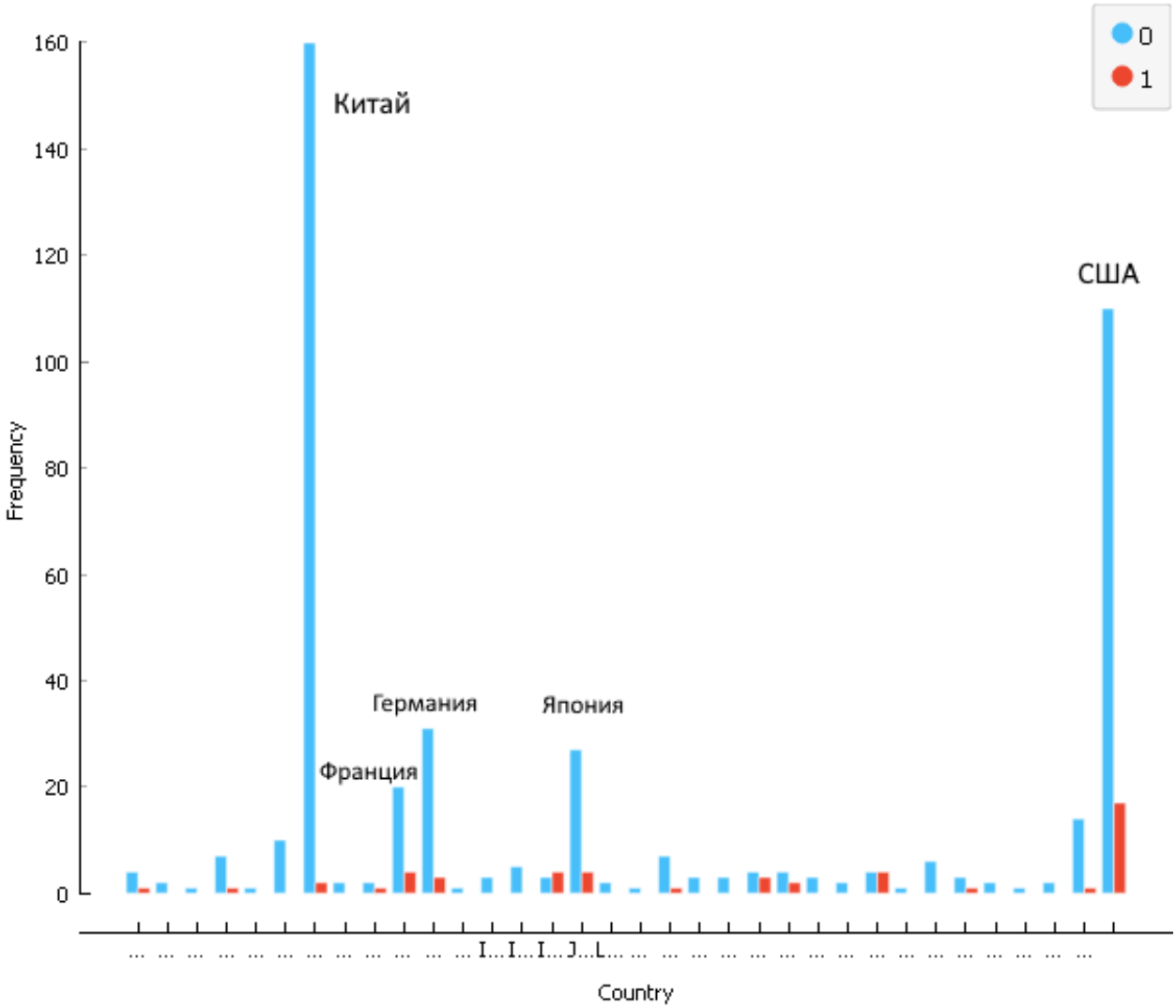

Рисунок 8: Множественная диаграмма распределения стран в top500, с разбиением на вхождение в 50 лучших. 1 - входит в 50 лучших; 0 - не входит.

Теперь вывод поменялся: *хоть и Китай имеет больше всех суперкомпьютеров, но среди пятерки лидеров он имеет самые слабые по производительности компьютеры. Очевидным лидером по производительности является США.*

Мы не рассматриваем другие страны, так как их влияние на общий вывод незначительно, для нас главное увидеть разницу в выводах, если с разной стороны смотреть на данные.

Это был пример того, как разнятся выводы анализа данных, если грамотно не изучить материал. В случае простого анализа - берется число определенных данных и делится на общее количество, однако в случае суперкомпьютеров так делать нельзя. Дело в том, что суперкомпьютеры тяжело исчислять в «штуках», так как кластеры в начале рейтинга значительно превосходят по параметру *flops* середину и конец рейтинга. К примеру, показатель первого кластера превосходит последний в практически в 640 раз! Такая разница говорит о том, что суперкомпьютеры следует анализировать, основываясь на их вычислительной мощности, а не количестве. В следующих разделах будем разделять данные на находящиеся в начале рейтинга и все остальные[7][8].

3.1.3 Поиск корреляций

В данном разделе представлены основные корреляции, наталкивающие на какой-либо вывод о суперкомпьютерах. Рассмотрим зависимость количества ядер в одном процессоре от общего количества ядер кластера:

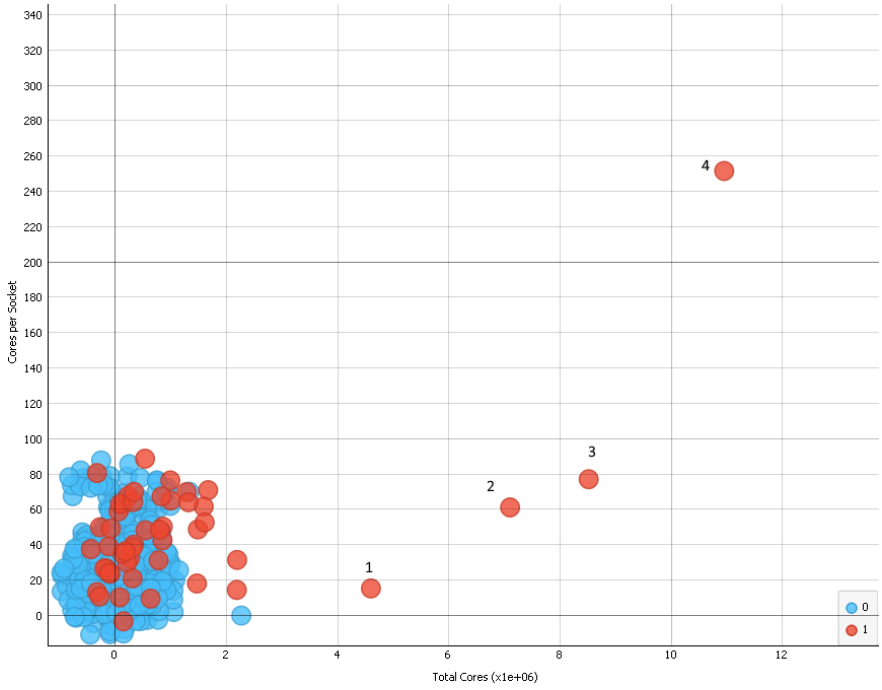

Рисунок 9: Зависимость количества ядер в одном процессоре от общего количества ядер. 1 - входит в 50 лучших в *top500*; 0 - не входит.

Анализируя данный график, можно сделать вывод, что нет какого-либо лидера в значении ядер в одном процессоре, суперкомпьютеры равномерно распределены по

данному параметру. Однако, можно сказать, что в основном самые производительные суперкомпьютеры используют 64-ядерные процессоры. Так же данный график имеет несколько «выбросов» - такие данные, которые не входят в основную массу данных, причем все из них входят в позицию 50 лучших кластеров. Точки 1-3 это просто лидеры списка с большим количеством процессоров - 1,2,10 место списка. Точка 4 - это суперкомпьютер «*Sunway Taihulight*», занимающий 7 место в рейтинге *top500*, и он имеет 260 ядер в одном процессоре! Производители данного кластера сами создавали специальный процессор для своего суперкомпьютера и сделали в нем 260 ядер. Такое решение обусловлено тем, что данный суперкомпьютер является многопрофильным, он предназначен для сложных расчётов, требуемых в производстве, медицине, добывающей промышленности, для прогнозирования погодных условий и анализа «больших данных»[9][10][13].

Теперь рассмотрим зависимость количества ядер от мощности кластера:

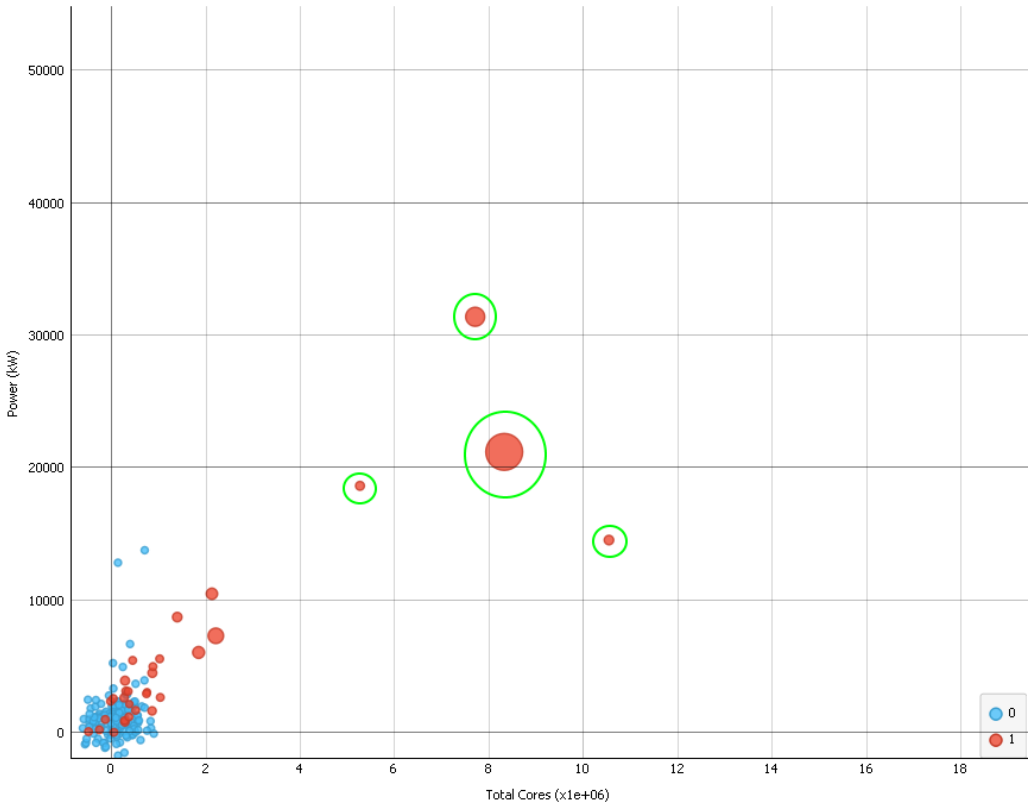

Рисунок 10: Зависимость количества ядер от мощности. 1 - входит в 50 лучших в *top500*; 0 - не входит.

На графике зеленым обведены выбросы из основной массы данных. Размер круга

прямо пропорционален параметру *Rmax*, самый большой круг - первое место в списке Top500. Проведя анализ графика можно заключить, что *зависимость количества ядер от мощности суперкомпьютера очевидна - прямая пропорциональность, однако можно добавить, что основная масса рейтинга Top500 имеет мощность порядка 5000 кВт и количество ядер около 50 тысяч, лидеры списка отличаются от остальных и находятся отдельно от других кластеров.*

3.1.4 RISC и CISC архитектуры

В мире суперкомпьютеров существует извечный вопрос: *Какая архитектура процессора лучше : RISC или CISC?* Архитектуры процессоров делятся на два основных вида RISC и CISC, впоследствии от них произошли производные архитектуры по типу ARM, RISC-V и т.д.[6] Попробуем проанализировать Top500 с этой точки зрения. Посмотрим на распределение архитектур в 50 лучших суперкомпьютеров, приняв что архитекутры ARM, RISC-V относятся к RISC:

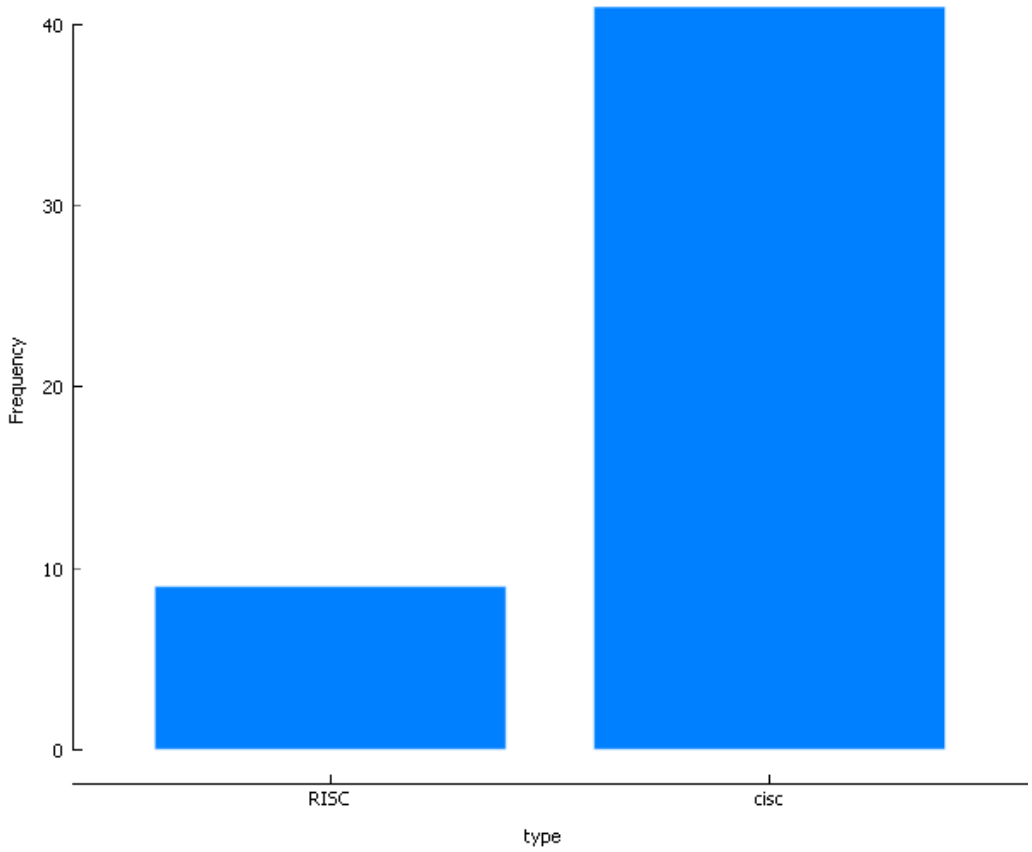

Рисунок 11: Диаграмма распределения архитектур

Видно существенное преобладание CISC архитектуры, однако рассмотрим

множественную диаграмму с делением, входит ли суперкомпьютер в десятку лучших, где 1 - входит, 0 - не входит:

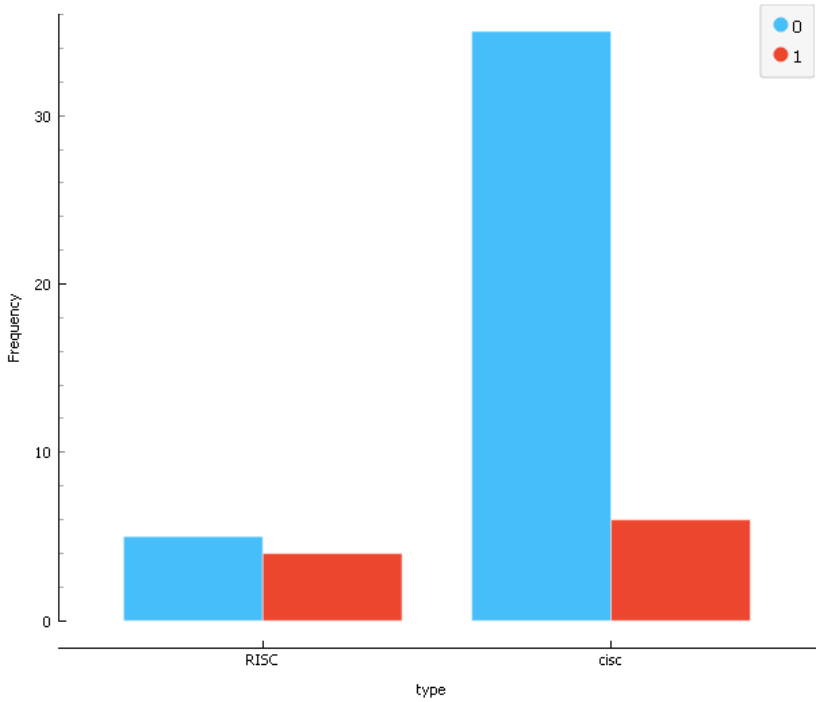

Рисунок 12: Множественная диаграмма распределения архитектур

Теперь вывод таков, что *в 50 лучших суперкомпьютеров преобладают CISC архитектуры, однако в десятки лучших архитектуры равны по количеству.* Для сравнения развития архитектур рассмотрим какая была ситуация в 2018 году:

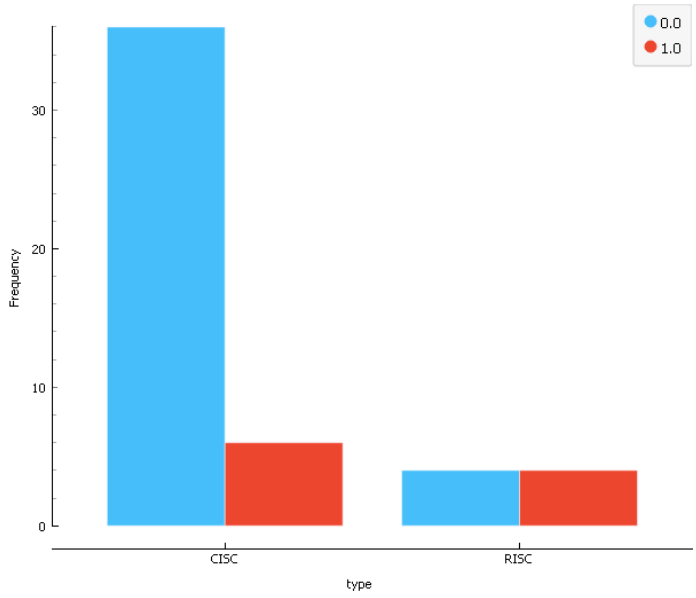

Рисунок 13: Диаграмма распределения архитектур в 2018 году

Как видно ситуация практически не изменилась, *на рынке суперкомпьютеров идет преобладание CISC архитектуры, основываясь на списке лучших 50 кластеров*.

Рассмотрим зависимость числа ядер от архитектуры:

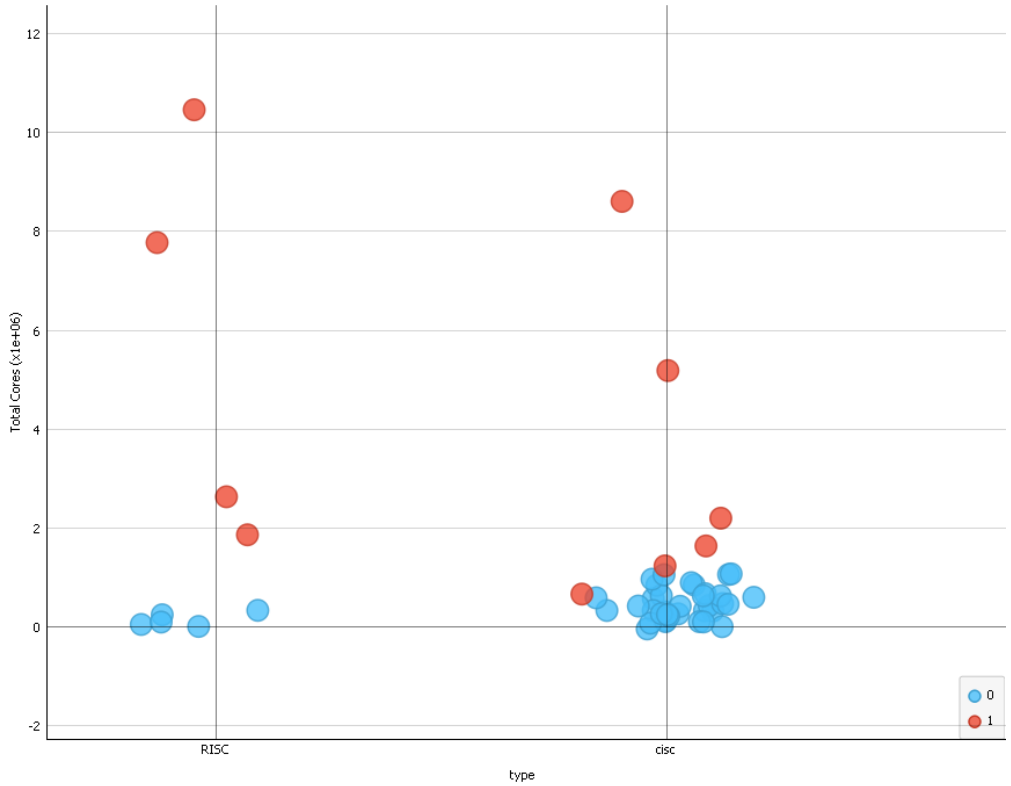

Рисунок 14: Распределение зависимости числа ядер от архитектуры

Распределение практически равное, если же посмотреть на 4 точки находящиеся на вершине распределения, то для них в среднем RISC архитектура имеет больше ядер, что может говорить о том, что *процессоры RISC-архитектуры более дорогие из-за большего количества ядер*, если не рассматривать влияние количества ядер в одном процессоре.

3.1.5 Подключение второй базы данных

Теперь к нашей основной базе данных добавим второстепенную рейтинг *Green500*, который определяет наилучшие суперкомпьютеры по энергоэффективности. В таком случае возникает вопрос: как распределены лучшие суперкомпьютеры мира в данном рейтинге? Посмотрим на диаграмму рассеяния ниже:

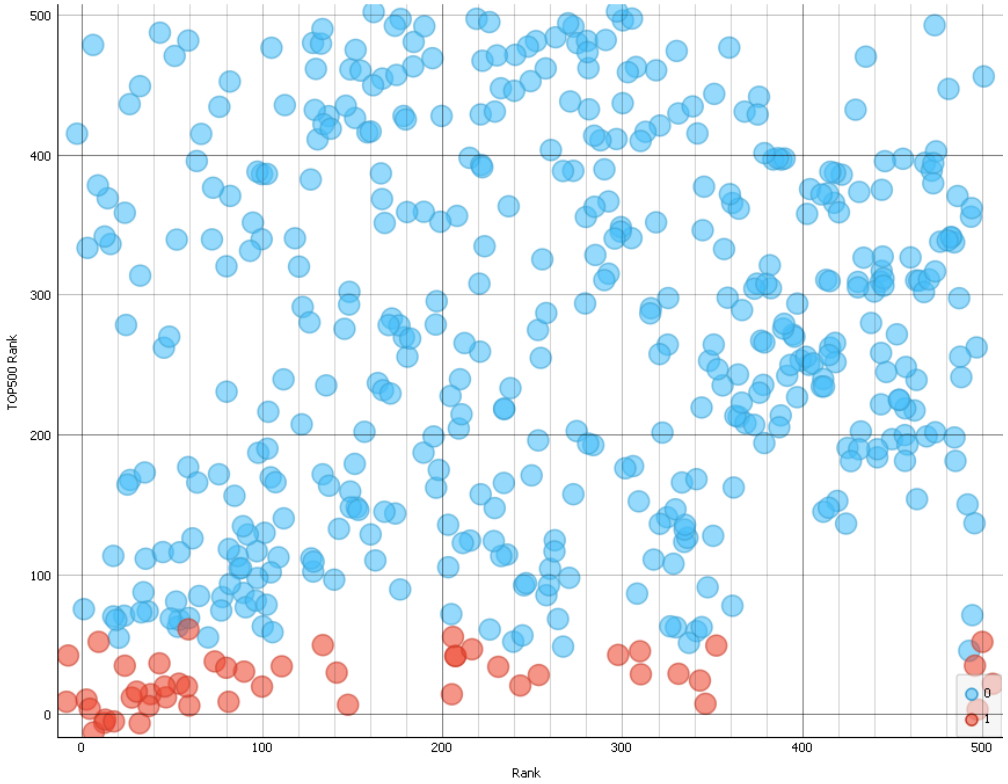

Рисунок 15: Распределение суперкомпьютеров в списке *Green500*. 1 - входит в 50 лучших в *top500*; 0 - не входит.

Из данного графика можно сделать вывод, что *лучший компьютер по производительности не означает, что он лучший по энергоэффективности*. Лучшие 50 кластеров равномерно распределены по всем позициям рейтинга *Green500*. Это говорит о том, что не все производители задумываются об энергоэффективности своего проекта, а просто вливают все ресурсы на то, чтобы улучшить производительность.

3.2 Анализ малой базы данных

Для решения задачи линейной регрессии возьмем для анализа малую базу данных - данные для предсказания цены вина. Известно, что цена вина менятся из года в год, производители хотят знать сколько будет стоить вино через 10,20,30 лет, чтобы узнать лучший момент для продажи - задача максимизации прибыли для анализа данных. В прошлом эта задача решалась с помощью профессионального дегустатора, однако такой способ очень субъективный и затратный. На сегодняшний день такую задачу решают методом линейной регрессии. Для работы возьмем данные исследования профессора Орли Ашенфельтера[12].

3.2.1 Визуализация представленных данных

Для анализа нам даны следующие данные : количество зимних осадков,

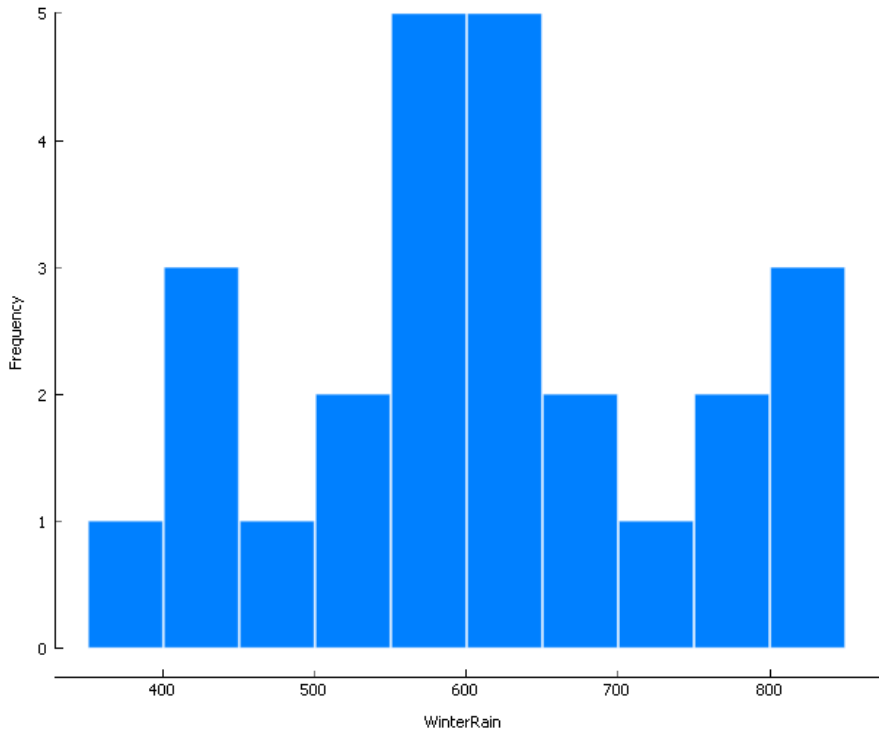

Рисунок 16: Диаграмма зимних осадков

количество летних осадков,

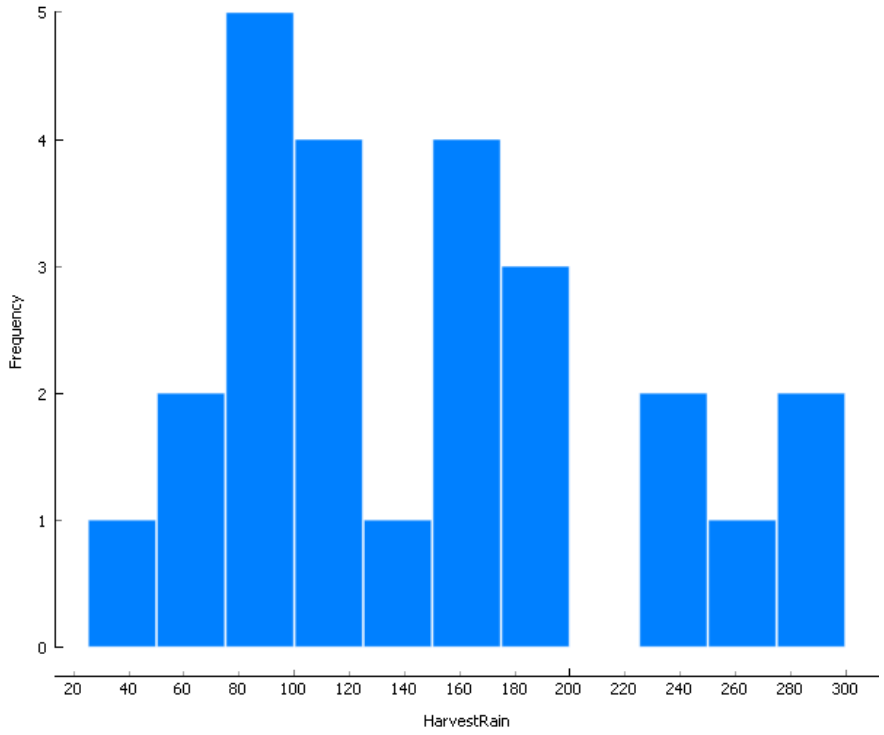

Рисунок 17: Диаграмма летних осадков

среднесезонная температура,

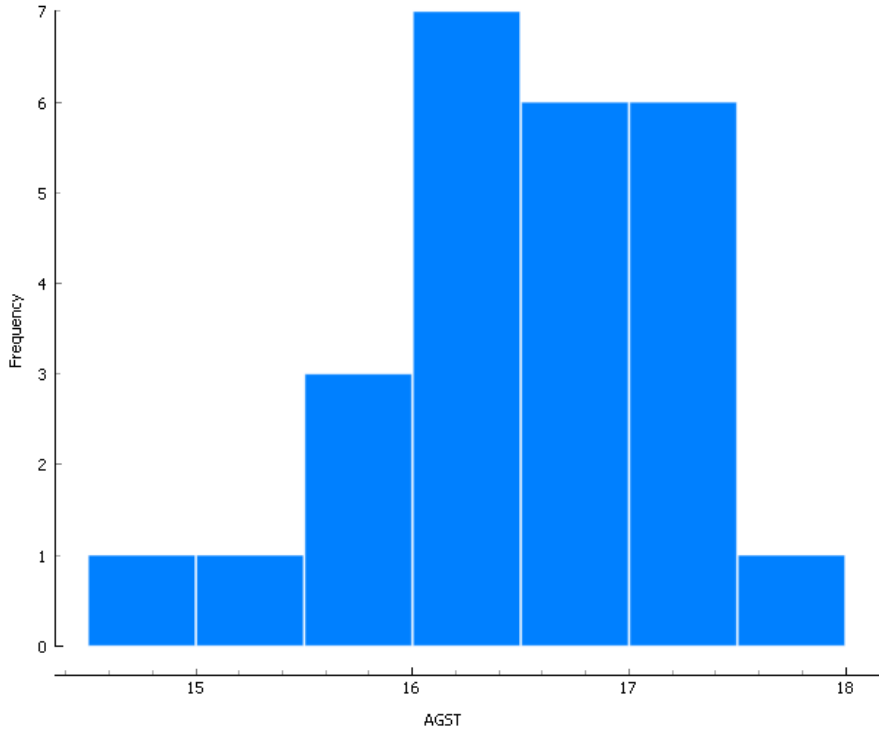

Рисунок 18: Диаграмма среднесезонной температуры

логарифм цены вина, логарифм используется для уменьшения влияния инфляции на исследование. На основе этой цены мы будем предсказывать будущие цены:

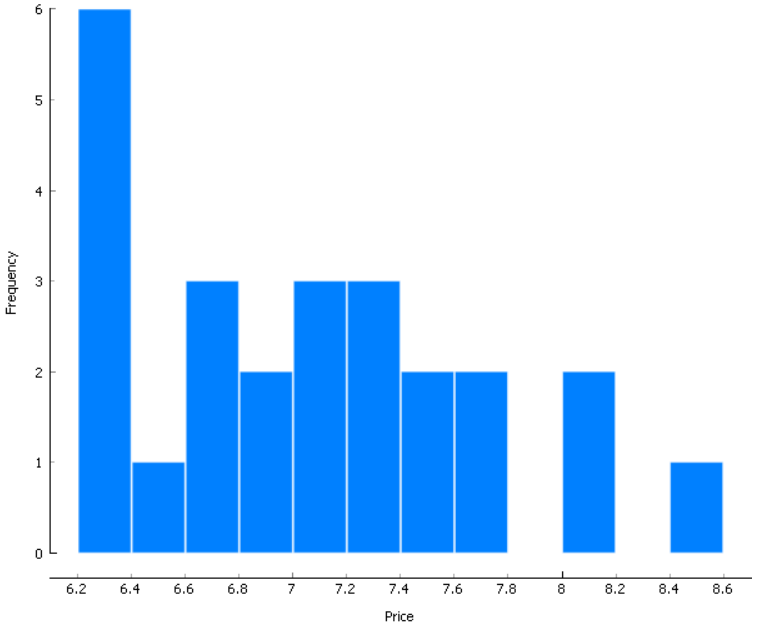

Рисунок 19: Диаграмма цены вина

3.2.2 Поиск корреляций

Для того, чтобы понимать как взаимодействуют наши данные между собой, построим несколько графиков. Зависимость цены и количества зимних осадков:

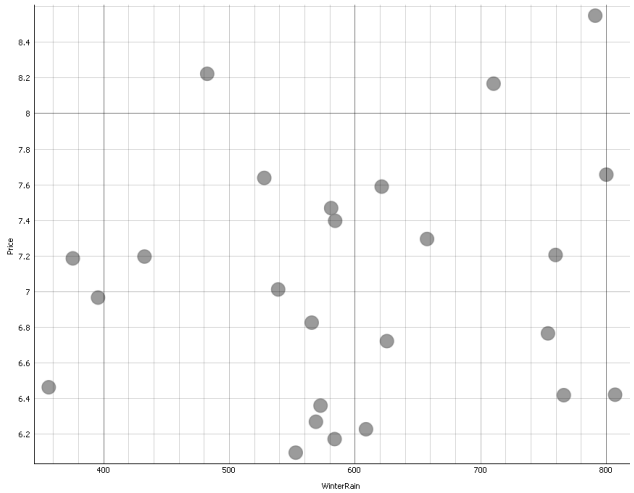

Рисунок 20: Распределение цены от зимних осадков

Прослеживается, что чем больше осадков зимой, тем в среднем цена на вино

будет выше. Просмотрим зависимость цены и количества летних осадков:

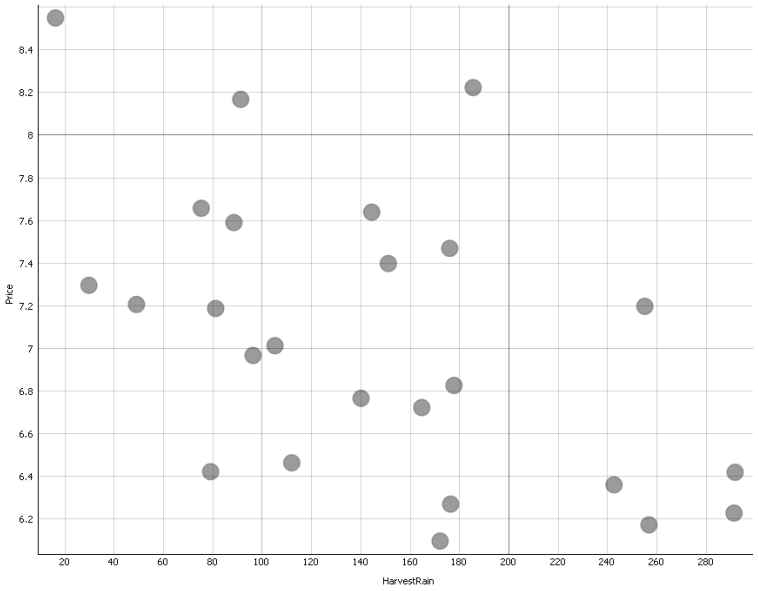

Рисунок 21: Зависимость цены от летних осадков

Есть четкая зависимость того, что чем меньше осадков летом, тем цена на вино выше. Теперь построим зависимость цены и среднесезонной температуры:

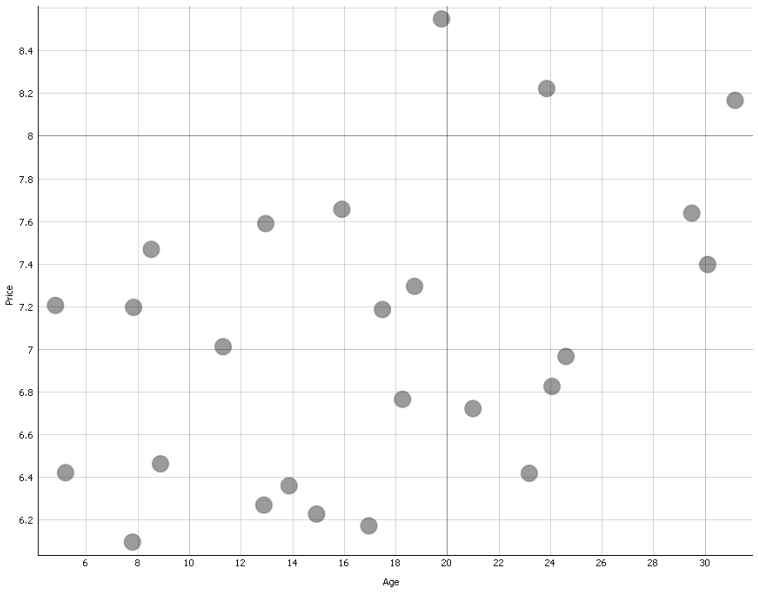

Рисунок 22: Распределение цены и среднесезонной температуры

Из данного рисунка можно сказать, что в холодные года вино будет дешевле, чем в тёплые. Наконец, посмотрим зависимость цены вина от его возраста:

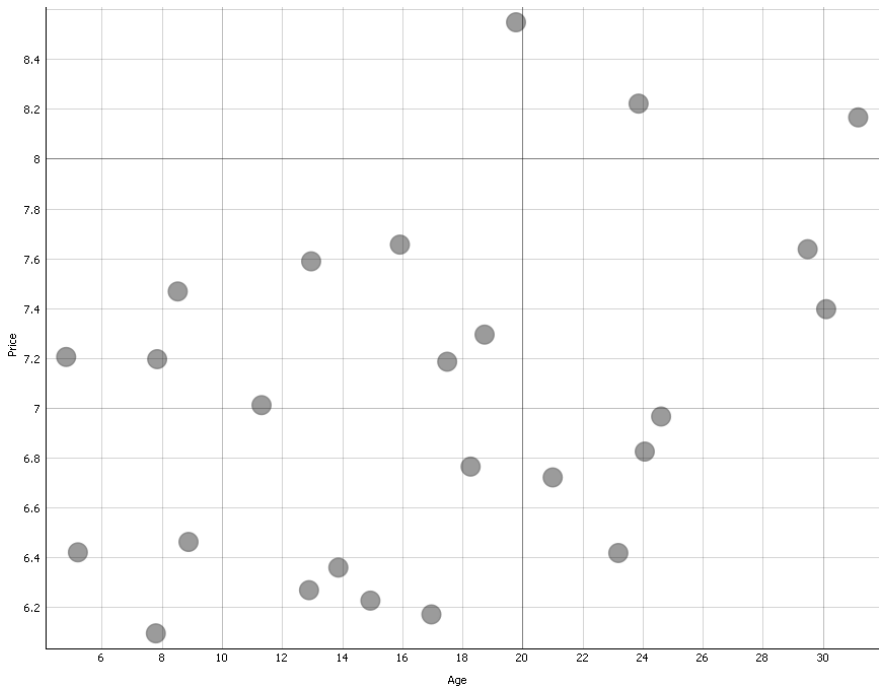

Рисунок 23: Зависимость цены вина от его возраста

Здесь нет четкой зависимости, но можно добавить, что среди вин старше 25 лет нет дешевых представителей.

Мы разобрались какие данные мы получили, и как они коррелируют между собой, теперь перейдем к решению задачи линейной регрессии.

3.2.3 Решение задачи линейной регрессии

Построив модель данных, анализированных выше, в результате мы получили следующие коэффициенты для уравнения линейной регрессии:

Имя	Коэффициент
Const	-3.42998
Зимние осадки	0.00107551
Среднесезонная температура	0.607209
Летние осадки	-0.00397153
Возраст	0.0239308

Коэффициенты получились согласно нашим корреляциям: где корреляция положительна - коэффициент > 0, где отрицательна - наоборот.

Итоговое уравнение выглядит следующим образом:

$$\ln P = -3.42998 + 0.00107551 \cdot WR + 0.607209 \cdot AGST - 0.00397153 \cdot SR + 0.0239308 \cdot AGE,$$

где P - цена вина, WR - количество зимних осадков, AGST- среднесезонная температура, SR - количество летних осадков, AGE - возраст вина.

Таким образом была решена задача линейной регрессии - получено уравнения для логарифма цены вина, которое можно использовать для предсказания будущих цен.

4 Выводы

Изучение основных методов анализа данных, позволило сделать следующие выводы.

Анализ данных — это быстроразвивающаяся отрасль в современном мире, которая совмещает в себе огромное количество методов обработки и визуализации данных. Анализ данных превращается в целое искусство, аналитик принимает решение какие методы лучше использовать для той или иной ситуации. Такое разнообразие выбора может отталкивать неопытных людей из-за незнания с чего начать работу с данными.

К **основным задачам** анализа данных можно отнести:

1. Сбор данных
2. Визуализация данных
3. Организация данных в пригодные для использования форматы
4. Поиска ответов в данных на конкретные вопросы
5. Выявление тенденций, в том числе негативных отклонений от плана, прогнозирование и получение рекомендаций.

В ходе анализа выявлены следующие **закономерности**: любой анализ должен начинаться с изучения рабочей области данных, в случае если неквалифицированный специалист будет проводить анализ узкой базы данных, то это может привести к плохим последствиям. В течение работы необходимо изучать связь между данными, как они коррелируют, для формирования правильных выводов. В ситуации, где необходимо предсказать какой-либо численный параметр может подойти метод линейной регрессии. Надежный и точный анализ данных требует только хорошего понимания и достаточной подготовки в области исследуемых данных, иначе результаты приведут нас к ложному решению.

На сегодняшний день данные обитают вокруг нас в каждой отрасли, крупные компании нуждаются в качественном аналитике данных для решения тех или иных задач. Результаты данного проекта способствуют дальнейшему исследованию более продвинутых методов анализа данных таких, как нейронные сети, метод случайного леса.

5 Список использованных источников

1. K. Kelley What is Data Analysis? Methods, Process and Types Explained / K. Kelley [Электронный ресурс] // Simple learn : [сайт]. — URL: https://www.simplilearn.com/data-analysis-methods-process-types-article (дата обращения: 24.12.2022).

2. И.Клейнер Онлайн-курс: Анализ данных: Просто и доступно / И.Клейнер [Электронный ресурс] // Stepik: Онлайн-курсы : [сайт]. — URL: https://stepik.org (дата обращения: 24.12.2022).

3. V. Kanade What Is Linear Regression? / V. Kanade [Электронный ресурс] // Spice works : [сайт]. — URL: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/ (дата обращения: 25.12.2022).

4. E. Strohmaier Supercomputing: What have we learned from the TOP500 Project? / E. Strohmaier, H.W. Meuer [Электронный ресурс] // Escholarship : [сайт]. — URL: https://escholarship.org/content/qt3212m88g/qt3212m88g.pdf?t=p0jv4hv=lg (дата обращения: 18.01.2023).

5. Top500 list / E. Strohmaier, M. Meuer, J. Dongarra, H. Simon [Электронный ресурс] // Рейтинг Top500 : [сайт]. — URL: https://www.top500.org (дата обращения: 26.12.2022).

6. E. Engheim Что означает RISC и CISC? / E. Engheim [Электронный ресурс] // Habr : [сайт]. — URL: https://habr.com/ru/company/selectel/blog/542074/ (дата обращения: 15.01.2023).

7. SUPERCOMPUTERS MARKET - GROWTH, TRENDS, COVID-19 IMPACT, AND FORECASTS (2023 - 2028) / [Электронный ресурс] // Mordorintelligence : [сайт]. — URL: https://www.mordorintelligence.com/industry-reports/supercomputer-market (дата обращения: 27.12.2023).

8. T. Trader Top500 Results: Latest List Trends and What's in Store / T. Trader [Электронный ресурс] // HPCwire : [сайт]. — URL: https://www.hpcwire.com/2017/06/19/49th-top500-list-announced-isc/ (дата обращения: 27.12.2022).

9. Тютляева Е.О., Одинцов И.О., Московский А.А., Мармузов Г.В. Тенденции развития вычислительных узлов современных суперкомпьютеров // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2019. Т. 8, № 3. С. 92–114. DOI: 10.14529/cmse190305

10. Awais K., Hyogi S., Sudharsha S. V. , Ali R. B. An Analysis of System Balance and Architectural Trends Based on Top500 Supercomputers // ResearchGate. 2021. DOI: 10.1145/3432261.3432263

11. Рынок анализа данных: инвестиции в бурно развивающуюся отрасль / [Электронный ресурс] // Газпром-инвестиции : [сайт]. — URL: https://gazprombank.investments/blog/market/data-analysis/ (дата обращения: 24.12.2022).

12. Case study: The Bordeaux equation / [Электронный ресурс] // Bookdown : [сайт]. — URL: https://bookdown.org/egarpor/PM-UC3M/lm-i-lab-wine.html (дата обращения: 17.01.2023).

13. Wikipedia / [Электронный ресурс] // Википедия : [сайт]. — URL: https://ru.wikipedia.org (дата обращения: 17.01.2023).

14. А. Семёнов Технологии Big Data / А. Семёнов [Электронный ресурс] // Uplab : [сайт]. — URL: https://www.uplab.ru/blog/big-data-technologies/ (дата обращения: 26.12.2022).