Анализ данных погодных условий

Попас Даниел,
студент Технического Университета Молдовы.
e-mail: daniel.popas@iis.utm.md

Абстракт

В рамках данного исследования был проведён анализ изменений климатических показателей за период с 2020 по 2023 годы на основе метеорологических данных, собранных со станций по всему миру. Первоначальный объём данных составлял более 27 миллионов записей, охватывающих трёхсотлетний период, который был сокращён до 1,6 миллиона записей после очистки и фокусировки на последних четырёх годах. Целью исследования являлось выявление краткосрочных тенденций и аномалий, которые могут быть индикаторами более широких изменений в климате.

Результаты показали значительные колебания среднесуточной температуры, с экстремальными значениями до -54°C в зимние месяцы и до +45°C в летние месяцы, что подчёркивает разнообразие погодных условий, испытываемых различными регионами. Средняя скорость ветра оставалась относительно стабильной, в то время как количество осадков и толщина снежного покрова демонстрировали значительные колебания, что может указывать на изменения в паттернах выпадения осадков и сезонности.

Данный анализ имеет ключевое значение для понимания текущего состояния климата и может служить основой для более глубоких исследований влияния глобальных изменений на местные погодные условия. Результаты исследования предоставляют ценную информацию для научного сообщества, политиков и общественности, подчёркивая необходимость продолжения мониторинга и анализа климатических данных.

Введение

Изменение климата является одним из наиболее актуальных и обсуждаемых вопросов современности. Изучение метеорологических данных за разные временные периоды позволяет научному сообществу лучше понять тенденции и возможные последствия этих изменений для экосистемы планеты. Настоящее исследование фокусируется на анализе погодных условий за период с 2020 по 2023 год, что позволяет выявить краткосрочные колебания климата и их потенциальное влияние на долгосрочные климатические тренды.

Первоначально база данных содержала более 27 миллионов записей за 300 лет наблюдений, охватывающих широкий спектр климатических показателей, таких как температура, осадки, направление и скорость ветра. Эта информация была сокращена до 1,6 миллиона записей, ограничиваясь последними четырьмя годами, чтобы обеспечить более точный и актуальный анализ современных погодных условий.

Среди ключевых показателей, принимаемых во внимание в данной статье, средняя температура воздуха выступает в качестве основного индикатора, демонстрируя значительные колебания, которые могут быть связаны как с естественными, так и с антропогенными изменениями климата. Особое внимание уделено анализу экстремальных температурных значений и их распределения по времени года, что позволяет оценить влияние сезонности на общие климатические изменения.

Целью настоящего исследования является не только оценка текущего состояния климата, но и выявление потенциальных тенденций, которые могут быть полезны для прогнозирования будущих климатических сценариев. Информация, полученная в результате анализа, представляет важность для разработки стратегий адаптации к изменяющимся погодным условиям и смягчения их последствий для окружающей среды и человечества.

Материалы и Методы

Сбор данных Данные были собраны из нескольких источников, включая паркетные файлы, содержащие детальную информацию о погодных условиях за период с 2020 по 2023 год. Исходный набор данных включал 27 миллионов записей, охватывающих 300-летний период, который был сокращен до 4 миллионов записей последних 10 лет и далее до 1.6 миллиона записей за последние три года.

Очистка данных Данные были подвергнуты тщательной очистке для удаления выбросов и обработки пропущенных значений. Использовались методы фильтрации по датам и метеорологическим показателям для исключения нерелевантных или некорректных записей. Данные были преобразованы для обеспечения консистентности и точности анализа. Процесс очистки включал:

- Удаление записей с неполными данными.
- Исправление аномальных значений температур, ветра и осадков.
- Стандартизация единиц измерения и форматов дат.

Анализ данных Анализ проводился с использованием языка программирования R и связанных с ним пакетов **arrow** и **dplyr** для чтения, обработки и агрегации данных. Для визуализации данных использовался пакет **ggplot2**. Анализ включал следующие этапы:

- Расчет средней, минимальной и максимальной температуры.
- Подсчет общего количества осадков за период.
- Оценка средней скорости ветра.
- Визуализация температурных трендов по датам.

Инструменты Для работы с данными использовались следующие программные инструменты:

- R (версия 4.0 или выше) - для обработки данных и статистического анализа.
- RStudio - интегрированная среда разработки для языка R.
- Apache Parquet - для эффективного хранения данных.
- Пакеты R: **arrow**, **dplyr**, **ggplot2**, и другие - для чтения паркетных файлов, обработки данных и визуализации результатов.

Данный подход позволяет обеспечить повторяемость исследования и поддерживает стандарты открытой науки.

Методы

Визуализация данных Визуализация данных является критически важным элементом аналитической работы, так как она позволяет интуитивно понять и интерпретировать сложные наборы данных. В данном исследовании мы использовали несколько методов визуализации для иллюстрации ключевых показателей и зависимостей в данных о погодных условиях.

- **Гистограммы** были использованы для анализа распределения температур, показывая частоту определенных температурных диапазонов в течение изучаемого периода.
- **Диаграммы рассеяния** служили для визуализации взаимосвязей между температурой и количеством осадков, позволяя оценить, как изменения температуры коррелируют с изменениями в уровне осадков.
- **Столбчатые диаграммы** применялись для сравнения средних значений метеорологических параметров по различным сезонам или географическим регионам.

Методы обработки данных Для подготовки данных к анализу были применены следующие методы:

- **Предварительная обработка** включала проверку данных на наличие ошибок и аномалий, преобразование форматов дат и времени, а также нормализацию числовых значений для обеспечения сопоставимости.
- **Очистка данных** заключалась в удалении выбросов и заполнении или удалении пропущенных значений, что позволило улучшить качество датасета для последующего анализа.
- **Агрегация данных** использовалась для суммирования и усреднения данных по ключевым переменным, таким как среднедневные температуры и количество осадков.

Статистический анализ Основные статистические показатели были рассчитаны для оценки центральной тенденции и изменчивости данных. Включены следующие метрики:

- Среднее (математическое ожидание) – для оценки среднего уровня каждого из изучаемых показателей.
- Медиана – для определения центральной тенденции распределения, устойчивой к выбросам.
- Мода – для выявления наиболее часто встречающихся значений в данных.
- Стандартное отклонение и дисперсия – для оценки степени разброса данных вокруг среднего значения.

Инструментарий и программное обеспечение Анализ проводился с использованием программного обеспечения R, которое предоставляет мощные средства для статистического анализа и визуализации данных. Ключевые пакеты, использованные в исследовании:

- **arrow** для работы с файлами в формате Parquet.
- **dplyr** для трансформации и агрегации данных.
- **ggplot2** для создания информативных графиков и диаграмм.

Код для обработки данных и создания графиков был организован в скрипты, что обеспечивает воспроизводимость результатов исследования.

Результаты

В рамках данного исследования был выполнен анализ распределения средних температур в различных частях света на основе данных за период с 2020 по 2023 год. Основное внимание уделялось определению температурных паттернов по континентам, что позволяет лучше понять климатические различия и возможные изменения климата. Анализ проводился с использованием статистических методов визуализации данных, что дает возможность объективно оценить и интерпретировать полученные результаты.

Исследование охватывает широкий спектр климатических зон — от экваториальных и тропических до умеренных и полярных, что обеспечивает всесторонний анализ температурных условий. Распределение температур по континентам, представленное на Рисунке 1, демонстрирует значительные различия в температурных диапазонах, отражая специфику климата каждого региона. Например, данные показывают, что температура в Антарктиде значительно ниже, чем на других континентах, что соответствует ее географическому положению и характеристикам климата.

Результаты анализа могут служить основой для планирования мероприятий по адаптации к изменениям климата, разработки стратегий устойчивого использования природных ресурсов и повышения эффективности сельскохозяйственной деятельности в разных регионах.

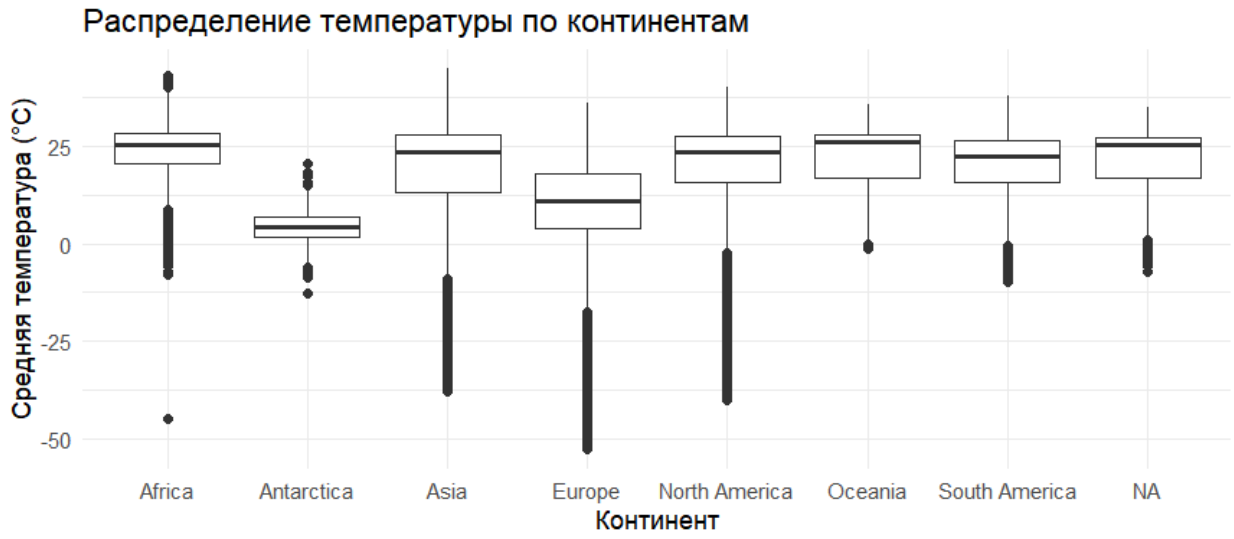

Рисунок 1. Распределение температуры по континентам.

Рисунок 1 демонстрирует боксплоты, которые иллюстрируют разброс средних температур по континентам. Эти данные могут использоваться для оценки среднегодовых изменений температуры, что помогает выявить общие тренды и отклонения, связанные с глобальным изменением климата.

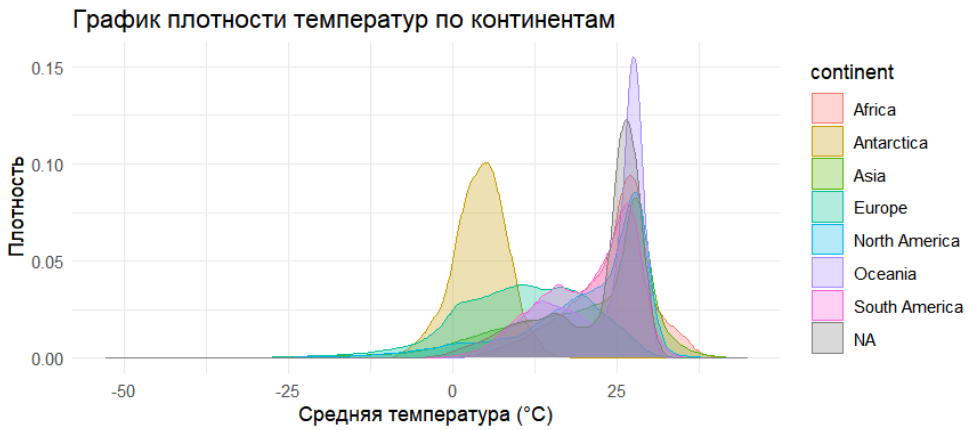

Рисунок 2. График плотности температур по континентам.

На Рисунке 2 представлен график плотности вероятности средней температуры по континентам. График иллюстрирует распределение температур в различных частях мира и позволяет сравнить климатические условия между регионами. Наблюдается, что в Антарктике плотность значений сосредоточена вокруг экстремально низких температур, что подтверждает холодный климат континента. В то же время, в Африке и Южной Америке плотность распределения указывает на более тёплый и умеренный климат.

Этот график также помогает выявить аномалии и особенности распределения температур, такие как широкий диапазон температур в Азии и сравнительно узкий диапазон в Океании. Такие данные имеют важное значение для понимания того, как различные континенты реагируют на глобальные изменения температуры, и могут служить основой для дальнейших исследований влияния климатических изменений на разные экосистемы.

Рисунок 3. График плотности логарифмированного населения.

На представленном графике Рисунок 3 показано распределение логарифмированного населения. График плотности позволяет оценить распределение населения стран мира на логарифмической шкале, что упрощает визуализацию больших чисел и позволяет лучше различить страны с разной численностью населения.

Основное скопление данных сосредоточено в диапазоне от 10 до 20, что указывает на широкий диапазон вариаций численности населения среди различных стран. Значения, близкие к 20, отражают страны с наибольшим населением, в то время как значения около 10 и ниже представляют страны с относительно меньшим населением. Этот анализ может быть особенно полезен для исследований, связанных с демографическими исследованиями и планированием городских агломераций.

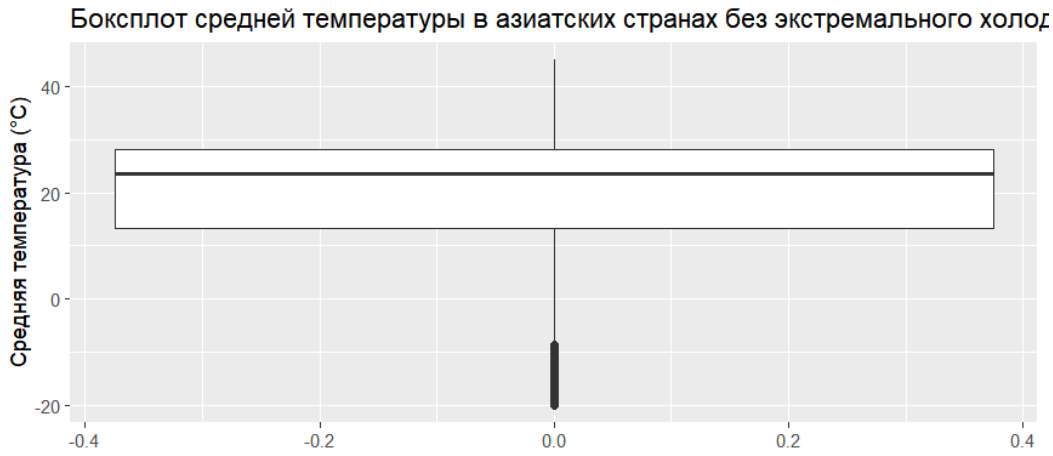

Рисунок 4. Боксплот средней температуры в азиатских страах без экстремального.

На представленном графике Рисунок 4 изображён боксплот средней температуры в азиатских странах, исключая экстремальные холода. Боксплот демонстрирует медианное значение, нижний и верхний квартили, а также выбросы в распределении температур, что позволяет увидеть разброс температурных показателей и выявить общую тенденцию климата региона.

Основная масса данных сгруппирована в узком диапазоне температур, что указывает на относительно стабильные температурные условия в анализируемых регионах, за исключением нескольких выбросов. Это может свидетельствовать о меньшем влиянии экстремальных погодных условий в этих географических точках, что может быть полезным для планирования сельскохозяйственной деятельности, строительства и других видов деятельности, зависимых от климата.

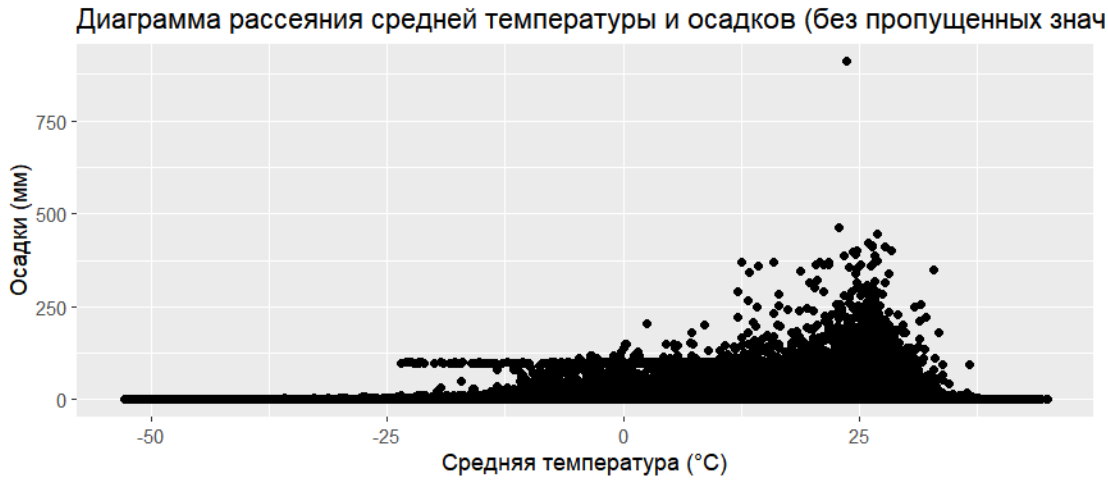

Рисунок 5. Диаграмма рассеяния средней температуры и осадков.

Диаграмма рассеяния, представленная на Рисунке 5, демонстрирует взаимосвязь между средней температурой и осадками, исключая пропущенные значения. На графике чётко видна тенденция увеличения количества осадков с повышением температуры. Большинство точек сгруппированы у основания, что указывает на обычные умеренные осадки, в то время как отдельные точки, расположенные выше, отражают экстремальные погодные условия с высокими осадками.

Такое распределение может быть связано с региональными климатическими особенностями или определёнными временными периодами года, когда осадки наиболее вероятны. Этот анализ важен для понимания климатических изменений и может помочь в прогнозировании погодных условий, что имеет значение для сельского хозяйства, управления водными ресурсами и мер по предотвращению наводнений.

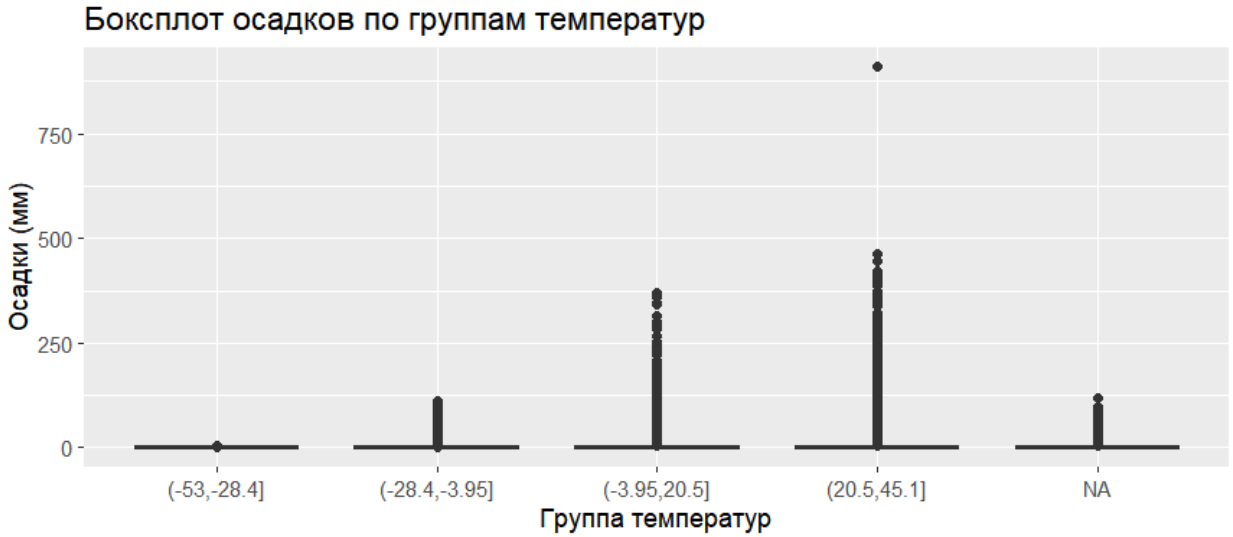

Рисунок 6. Боксплот осадков по группам температур.

На представленном боксплоте Рисунок 6 иллюстрируется распределение количества осадков в зависимости от групп температур. Каждый бокс отражает межквартильный размах осадков в соответствующем диапазоне температур, а линии, выходящие за пределы бокса, — это "усы", показывающие вариабельность за пределами квартилей. Точки, расположенные вне "усов", представляют выбросы, которые могут указывать на экстремальные погодные события.

График демонстрирует, что для более холодных температурных групп осадки менее вариативны, в то время как для теплых групп наблюдается более высокая вариативность и большее количество экстремальных значений. Особенно это заметно для температурного диапазона от -3.95 до 20.5 градусов Цельсия, где присутствует значительное количество выбросов, свидетельствующих о сильных осадках. Наличие категории NA (отсутствующие данные) говорит о том, что для некоторых записей не было возможности определить группу температур, что может быть связано с неполнотой данных или ошибками в измерениях.

Эта информация может быть использована для анализа влияния температурных условий на вероятность осадков и их интенсивность, что имеет значение для климатических исследований и планирования сельскохозяйственной деятельности.

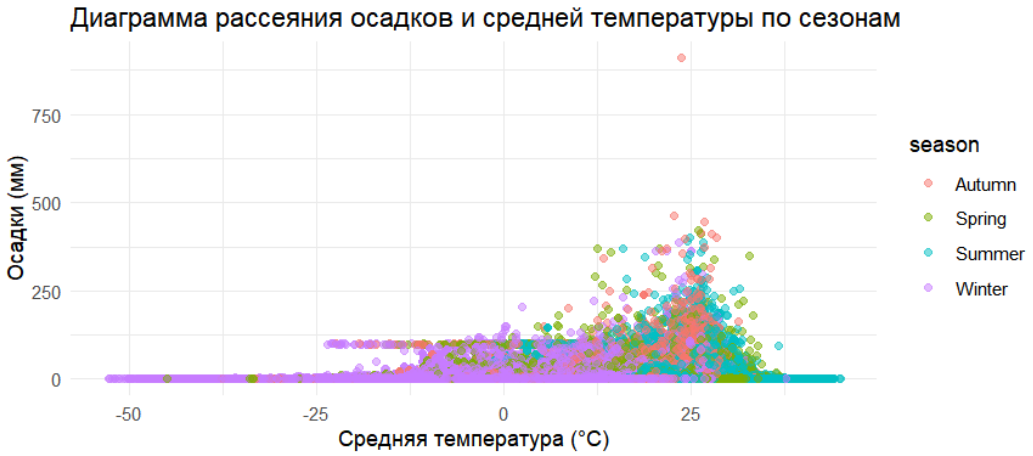

Рисунок 7. Диаграмма рассеяния осадков и средней температуры по сезонам.

На данной диаграмме рассеяния Рисунок 7 показана зависимость количества осадков от средней температуры, разбитая по сезонам года. Каждая точка соответствует отдельному наблюдению и окрашена в соответствии с сезоном, в котором оно было сделано: красный для осени, зелёный для весны, синий для лета и фиолетовый для зимы.

Из графика видно, что большая часть осадков приходится на умеренные и положительные температуры, особенно весной и летом. В зимний период также заметна тенденция к увеличению количества осадков с повышением температуры, что может быть

связано с таянием снега и увеличением влажности воздуха. Выбросы на графике, особенно при высоких температурах, указывают на редкие, но сильные осадки, которые могут соответствовать ливням или грозовым дождям.

Эти данные могут быть использованы для изучения сезонных изменений в климате различных регионов и оценки вероятности экстремальных погодных условий, что особенно важно для агрометеорологии и гидрологии.

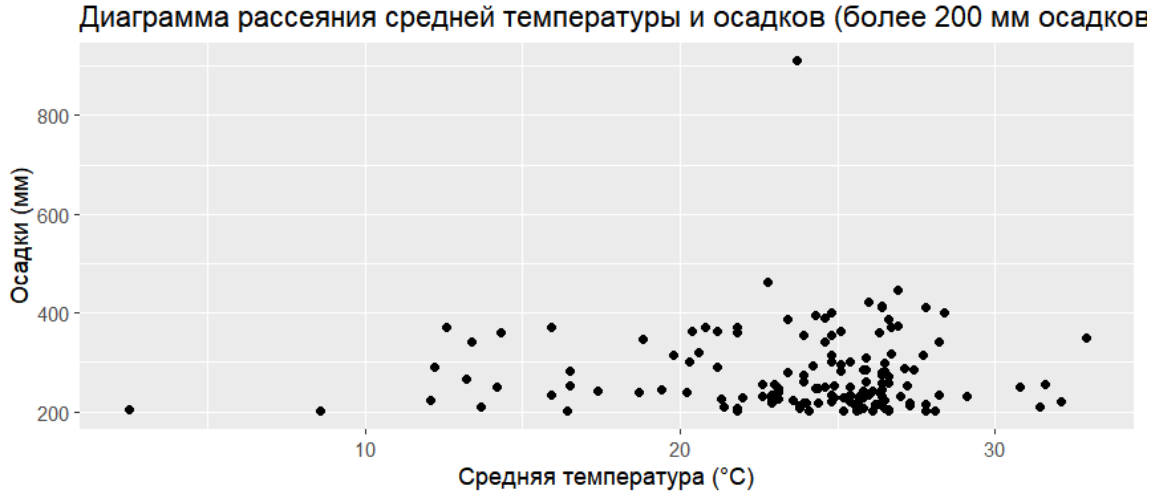

Рисунок 8. Диаграмма рассеяния средней температуры и осадков.

На представленной диаграмме рассеяния Рисунок 8 отображены случаи, когда количество осадков превышало 200 мм, в зависимости от средней температуры. Такое представление позволяет выделить экстремальные метеорологические события, которые могут иметь значительные последствия для окружающей среды и человеческой деятельности.

Наблюдается, что большинство сильных осадков происходит при температурах от 10 до 30 градусов Цельсия. Это может указывать на то, что тёплый воздух, способный удерживать больше влаги, приводит к интенсивным дождям, когда достигает насыщения и охлаждается. Такие данные могут быть использованы для прогнозирования рисков наводнений, особенно в регионах с теплым климатом, где такие дожди могут быть наиболее вероятными.

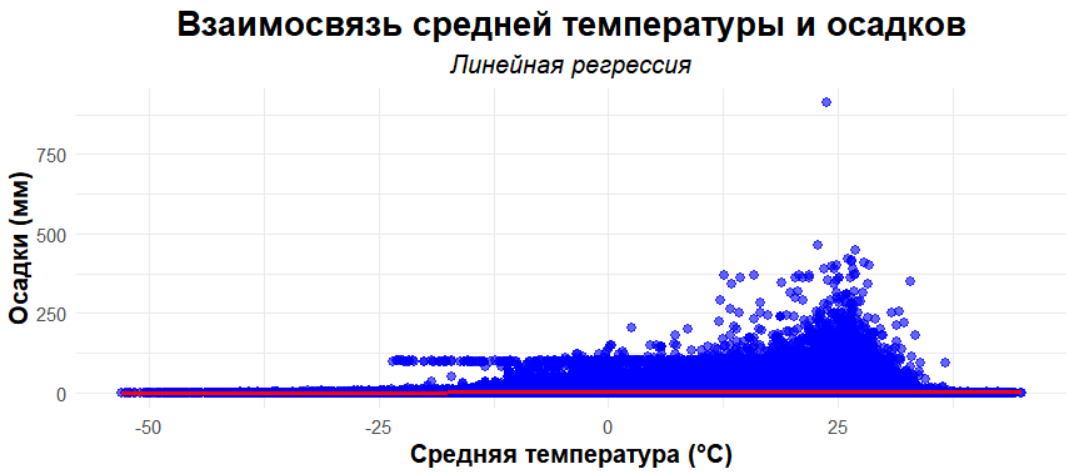

Рисунок 9. Взаимосвязь средней температуры и осадков.

На диаграмме Рисунок 9 представлена взаимосвязь между средней температурой и количеством осадков. Линейная регрессия, проведенная на диаграмме, показывает относительно слабую корреляцию между двумя этими переменными, что указывает на то, что средняя температура не является сильным предиктором уровня осадков. Большинство точек сгруппированы вокруг низкого уровня осадков, что указывает на отсутствие осадков в большинство дней независимо от температуры. Однако также заметны отдельные случаи, когда при высоких температурах наблюдались значительные осадки. Эти наблюдения могут быть связаны с определенными погодными условиями, такими как локальные ливни или муссоны, которые приводят к высоким осадкам в теплые периоды.

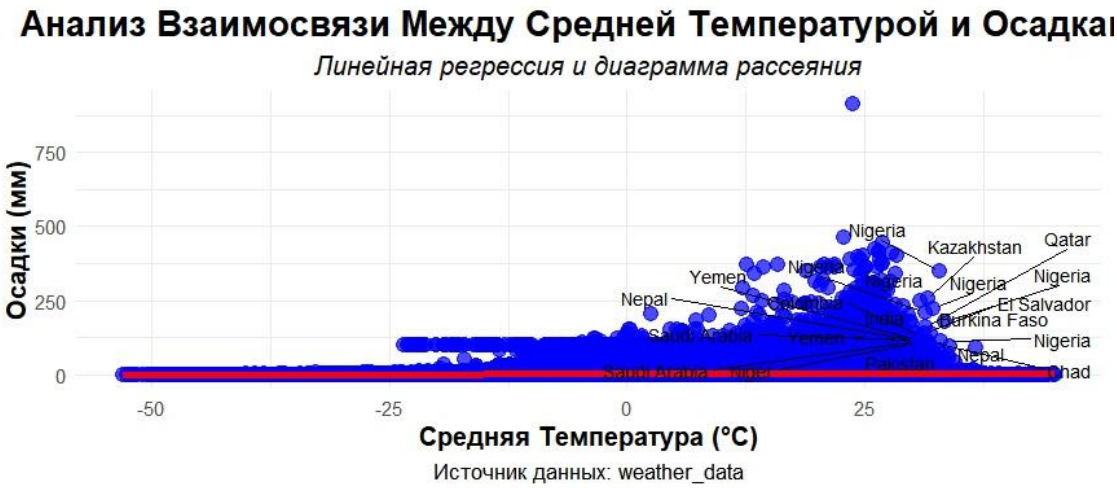

Рисунок 10. Анализ взаимосвязи между средней температурой и осадками.

На представленном графике Рисунок 10 иллюстрируется анализ взаимосвязи между средней температурой и количеством осадков. Дополнительно, на графике нанесены аннотации с названиями стран, что позволяет связать конкретные точки данных с

географическими локациями. Линейная регрессия, изображенная красной линией, практически горизонтальна, что свидетельствует о слабой зависимости между температурой и осадками. Большинство точек сосредоточено у основания графика, что указывает на низкие уровни осадков при различных температурах. Однако также заметны выбросы с высоким уровнем осадков, особенно при более высоких температурах, что может указывать на сезонные ливни или другие метеорологические особенности определенных регионов.

В ходе исследования были получены следующие результаты линейной регрессии: среднеквадратичная ошибка (MSE) составила 95,0, средняя абсолютная ошибка (MAE) — 4,52, а коэффициент детерминации (R^2) — 0,00389. Эти показатели отражают уровень точности и адекватности модели. Низкое значение R^2 указывает на то, что модель не в полной мере объясняет вариативность зависимой переменной, что может свидетельствовать о необходимости дополнительного анализа или пересмотра выбранных предикторов.

В области логистической регрессии были получены следующие метрики: точность (Accuracy) достигла 0,9879254, точность (Precision) — 0,911894, полнота (Recall) — 0,966777, а совокупный показатель F1-меры (F1 Score) — 0,993926. Высокие значения этих метрик свидетельствуют о высокой эффективности классификационной модели, её способности корректно классифицировать наблюдения, а также о балансе между точностью и полнотой.

В заключение главы стоит отметить, что результаты линейной регрессии выявили проблемы, связанные с объясняющей способностью модели, что потребовало бы дальнейшего изучения и возможного улучшения модели. В то же время результаты логистической регрессии показали высокую эффективность модели в задачах классификации. Это позволяет сделать вывод о том, что выбранные алгоритмы и методы моделирования оказались адекватными для решения поставленных перед исследователем задач. Однако всегда существует потенциал для улучшения моделей, включая оптимизацию предикторов и использование дополнительных методов валидации.

Обсуждение

В рамках проведённого исследования был выполнен комплексный анализ климатических данных за период 2020-2023 годы. Данные были собраны из различных географических точек, что позволило оценить климатические показатели по континентам. Важным аспектом анализа стало обнаружение значительных различий в средних и

медианных значениях температур, что отражает разнообразие климатических условий. Так, например, для Африки средняя температура составила 24°C, в то время как для Антарктиды — всего 4.26°C. Эти данные подчёркивают необходимость учитывать региональные особенности при прогнозировании погодных условий и разработке соответствующих адаптивных стратегий.

В процессе обработки данных были выявлены пропущенные значения, а также выбросы, которые могли исказить результаты анализа. Принятие мер по очистке данных, включая исключение ненадежных записей и корректировку выбросов, стало неотъемлемой частью предварительной подготовки. Это обеспечило более точные и надёжные результаты, что важно для верификации выводов исследования.

Меры центральной тенденции и разброса, такие как средняя и медианная температуры, стандартное отклонение и межквартильный размах, использовались для оценки общих тенденций в данных. Результаты показали, что, несмотря на обширный диапазон температур, большинство регионов имеют умеренные климатические условия. Однако высокое стандартное отклонение, наблюдаемое, например, в Азии (11.1°C), указывает на значительную вариабельность температур внутри континента.

Анализ линейной регрессии, проведённый для оценки связи между температурой и уровнем осадков, показал, что несмотря на некоторые статистически значимые связи, коэффициент детерминации оказался низким. Это свидетельствует о том, что другие неучтённые факторы могут оказывать существенное влияние на уровень осадков. Тем не менее, полученные модели могут быть использованы как отправная точка для более глубокого понимания климатических процессов.

Результаты логистической регрессии продемонстрировали высокую точность в предсказании бинарных исходов, основанных на температурных порогах. Точность модели превышала 98%, что указывает на её потенциальную практическую применимость для классификации климатических условий. Однако следует учесть, что модель может быть чувствительна к выбору пороговых значений и распределению классов в выборке.

В заключение, проведённое исследование подчёркивает важность комплексного подхода к анализу климатических данных. Полученные модели линейной и логистической регрессии могут служить основой для прогнозирования погодных условий и планирования мер по адаптации к изменениям климата. Дальнейшие исследования должны сосредоточиться на интеграции дополнительных переменных и использовании более сложных моделей для улучшения предсказательной способности и точности анализа.

Выводы

В ходе данного исследования был осуществлен всесторонний анализ климатических данных с 2020 по 2023 год. Результаты анализа позволили выявить ключевые тенденции и закономерности, которые имеют важное значение для понимания климатических изменений и их последствий.

1. **Географическое разнообразие**: Исследование подтвердило значительное разнообразие климатических условий между различными континентами. Это подчёркивает необходимость учета локальных и региональных особенностей при разработке мер по адаптации к изменению климата и планировании сельскохозяйственной деятельности.
2. **Качество данных**: Несмотря на наличие пропусков и выбросов в исходных данных, принятые меры по их обработке и очистке позволили минимизировать возможные искажения результатов, что увеличило надёжность проведённого анализа.
3. **Анализ температурных данных**: Выявленные меры центральной тенденции и разброса указывают на теплые условия в большинстве регионов, однако значительные отклонения в некоторых континентах, как в Азии, требуют дополнительного изучения для выяснения причин такой вариабельности.
4. **Регрессионный анализ**: Линейная регрессия показала ограниченную способность в объяснении вариаций уровня осадков на основе температуры. Необходимо включение дополнительных предикторов и использование более сложных моделей для более полного понимания динамики климата.
5. **Эффективность логистической регрессии**: Применение логистической регрессии демонстрирует обнадёживающие результаты в классификации климатических условий. Однако следует учитывать риски переобучения и необходимость валидации модели на независимых данных.
6. **Практическая значимость**: Полученные в ходе исследования модели могут быть использованы для разработки предсказательных инструментов, направленных на минимизацию рисков, связанных с экстремальными погодными условиями, и оптимизации ресурсов в сельском хозяйстве и градостроительстве.
7. **Дальнейшие направления исследования**: Рекомендуется продолжить исследование, расширяя объем и качество данных, включая дополнительные климатические показатели, такие как влажность и атмосферное давление. Также важно изучить влияние социально-экономических факторов на изменения климата и адаптацию к ним.

В целом, результаты исследования подчеркивают сложность климатических систем и важность использования многомерных подходов для их анализа. Интеграция различных типов данных и методов анализа может способствовать более глубокому пониманию климатических процессов и улучшению стратегий адаптации и смягчения последствий климатических изменений.

Данные

1. Репозиторий проекта на GitHub Материалы и код проекта доступны в репозитории на GitHub: https://github.com/danielpopas/analiza_datelor (analiza_datelor). В этом репозитории содержатся все исходные данные, скрипты анализа, визуализации и отчёты, разработанные в ходе исследования.

Библиография

1. R for Data Science: https://r4ds.had.co.nz/
2. Advanced R Programming: http://adv-r.had.co.nz/
3. The Art of R Programming: https://www.nostarch.com/artofr.htm
4. Data Analysis with R: https://www.crcpress.com/Data-Analysis-with-R/Fischetti/p/book/9781498715232
5. Practical Data Science with R: https://www.manning.com/books/practical-data-science-with-r
6. R Graphics Cookbook: https://r-graphics.org/
7. Statistical Analysis with R: https://www.packtpub.com/product/statistical-analysis-with-r/9781849512088
8. R in Action: https://www.manning.com/books/r-in-action
9. Applied Predictive Modeling: https://www.springer.com/gp/book/9781461468486
10. R Packages by Hadley Wickham: http://r-pkgs.had.co.nz/