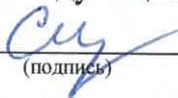




МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт информационных технологий (ИИТ)
Кафедра прикладной математики (ПМ)

Утверждаю
Заведующий кафедрой ПМ

(подпись) Смоленцева Т.Е.
«22» февраля 2025 г.

ЗАДАНИЕ

на выполнение курсовой работы

по дисциплине «Языки программирования для статистической обработки данных»

Студент Снигаренко Антон Владимирович

Группа ИНБО-22-23

Тема «Использование методов иерархической кластеризации и алгоритма k-means для определения факторов влияющих на объем выработки электроэнергии газовой турбиной»

Исходные данные: собранный студентом набор данных по теме работы

Перечень вопросов, подлежащих разработке, и обязательного графического материала:

Характеристика изучаемой предметной области, алгоритма, набора данных (описание текущего состояния исследуемой предметной области, выделение перспективных направлений исследований, применимость алгоритмов анализа и обработки данных, описание полей набора данных)

Математическая формулировка предлагаемого метода анализа и обработки данных (классическая постановка задачи, формулировка задачи статистической обработки данных, описание параметров, описание критерия качества решения конечной задачи)

Анализ полученной выборки данных с использованием предложенных методов анализа и обработки данных (описание последовательности действий или сценария обработки данных)

Построение визуализаций и качественных выводов по проделанной работе

Срок представления к защите курсовой работы:

до «23» мая 2025 г.

Задание на курсовую работу выдал



Подпись руководителя

Трушин С.М.

(ФИО руководителя)

«22» февраля 2025 г.

Задание на курсовую работу получил


Подпись обучающегося

Снигаренко А.В.

(ФИО обучающегося)

«22» февраля 2025 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	5
1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ	7
1.1 Описание предметной области.....	7
1.1.1 Значимость анализа данных в рассматриваемой области	7
1.1.2 Основные типы данных	8
1.1.3 Ключевые задачи статистического анализа	9
1.2 Основные методы статистической обработки данных.....	10
1.2.1 Описательная статистика.....	10
1.2.2 Методы проверки гипотез	12
1.2.3 Корреляционный анализ и регрессионное моделирование...	13
1.2.4 Методы кластеризации и классификации	16
1.2.5 Методы нейросетевых моделей	18
1.2.6 Примеры практического применения статистических методов	20
1.3 Язык программирования R и его возможности для статистического анализа	22
1.3.1 Основные преимущества языка R	22
1.3.2 Обзор ключевых библиотек.....	23
1.3.3 Примеры использования R в обработке данных	25
1.4 Программное обеспечение Glarus BI	26
1.4.1 Общая характеристика BI-системы	26
1.4.2 Основные функциональные возможности.....	28
1.4.3 Примеры использования Glarus BI для анализа данных	29

1.5	Сравнение возможностей R и Glarus BI	31
1.5.1	Преимущества R.....	31
1.5.2	Преимущества Glarus BI.....	32
1.5.3	Возможность совместного использования инструментов.....	34
2.	ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ ИССЛЕДОВАНИЯ.....	36
2.1	Описание используемых данных	36
2.1.1	Источники данных	37
2.1.2	Структура и характеристики данных	39
2.1.3	Предварительная обработка данных	39
2.2	Исследовательский анализ данных	40
2.2.1	Визуализация распределений.....	41
2.2.2	Корреляционный анализ.....	43
2.2.3	Выявление выбросов и трендов.....	44
2.3	Применение методов статистического анализа	44
2.3.1	Описательная статистика.....	45
2.3.2	Проверка гипотез.....	46
2.3.3	Регрессионный анализ	46
2.4	Машинное обучение в анализе данных	46
2.4.1	Классификация данных	46
2.4.2	Кластеризация	46
2.4.3	Нейросетевые модели	Ошибка! Закладка не определена.
2.5	Визуализация данных.....	48
2.5.1	Визуализация в R.....	48
2.5.2	Интерактивные дашборды в Glarus BI	48
2.5.3	Сравнение методов визуализации	48

3. АВТОМАТИЗАЦИЯ И ОТЧЁТНОСТЬ В АНАЛИЗЕ ДАННЫХ ..	49
3.1 Генерация отчётов в R	50
3.1.1 Обоснование необходимости автоматизации отчётов	50
3.1.2 Использование RMarkdown для создания отчетов.....	50
3.1.3 Экспорт отчетов в PDF, HTML, Word.....	50
3.2 Формирование интерактивных отчётов в Glarus BI	50
3.2.1 Различие между статичными и интерактивными отчетами..	50
3.2.2 Создание дашбордов в Glarus BI	50
3.2.3 Экспорт отчетов в Glarus BI	50
3.3 Сравнение инструментов R и Glarus BI.....	50
3.3.1 Анализ сильных и слабых сторон инструментов.....	50
3.3.2 Возможности интеграции R и Glarus BI	51
3.3.3 Применимость инструментов для различных типов задач ...	51
ЗАКЛЮЧЕНИЕ.....	52
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	53
ПРИЛОЖЕНИЕ А	55

Отчет об анализе брака и развода **Ошибка! Закладка не определена.**

ВВЕДЕНИЕ

В условиях постоянно растущего спроса на электроэнергию и ужесточения требований к эффективности и экологичности энергетических установок, оптимизация работы газотурбинных установок (ГТУ) приобретает особое значение. Газовые турбины являются одним из ключевых элементов современной энергетики, обеспечивая как базовую, так и пиковую генерацию электроэнергии. Объем вырабатываемой ими электроэнергии зависит от сложного взаимодействия множества эксплуатационных параметров (таких как температура на входе и выходе, давление, скорость вращения) и внешних условий (например, температура окружающей среды, влажность). Выявление и понимание наиболее значимых факторов, влияющих на производительность ГТУ, является критически важной задачей для повышения их эффективности, надежности и снижения эксплуатационных затрат.

Традиционные методы анализа зачастую не позволяют в полной мере учесть многомерность и нелинейность связей между различными параметрами работы ГТУ. В связи с этим, применение современных методов интеллектуального анализа данных, в частности алгоритмов кластеризации, открывает новые возможности для исследования больших объемов эксплуатационных данных. Методы иерархической кластеризации и алгоритм k -средних (k -means) относятся к классу алгоритмов неконтролируемого обучения и позволяют выявлять естественные группы или кластеры в данных без предварительной разметки.

Целью данной работы является применение методов иерархической кластеризации и алгоритма k -means для анализа эксплуатационных данных газовой турбины с целью идентификации различных режимов ее работы и определения ключевых факторов, оказывающих статистически значимое влияние на объем выработки электроэнергии. Предполагается, что такой подход позволит выявить скрытые закономерности и взаимосвязи, которые

могут быть использованы для оптимизации управления ГТУ, прогнозирования ее состояния и повышения общей эффективности энергопроизводства.

В данной курсовой работе будет использоваться синтетический dataset, с которым будут выполнены задачи курсовой работы.

1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ

1.1 Описание предметной области

1.1.1 Значимость анализа данных в рассматриваемой области

В рамках исследования рассматривается применение передовых методов анализа данных, в частности иерархической кластеризации и алгоритма k-средних, для идентификации и оценки значимости факторов, оказывающих влияние на объем генерации электрической энергии газовыми турбинами. Эффективность эксплуатации газотурбинных установок (ГТУ) напрямую зависит от множества операционных и внешних параметров. Целью работы является выявление скрытых закономерностей в многомерных массивах данных, регистрируемых в процессе работы ГТУ, с последующим формированием кластеров, характеризующих различные режимы работы и соответствующие им уровни выработки.

Иерархическая кластеризация будет использована для первоначального exploratory анализа, позволяющего определить естественную структуру данных и оптимальное количество кластеров без априорных предположений. Далее, алгоритм k-средних будет применен для более четкой сегментации данных на основе выявленного числа кластеров, что обеспечит формирование групп наблюдений со схожими характеристиками. Анализ центроидов полученных кластеров и распределения ключевых переменных (таких как температура окружающей среды, давление на входе, влажность, состав и расход топлива, параметры вибрации, температура выхлопных газов и т.д.) внутри каждого кластера позволит выявить те параметры, которые наиболее

сильно коррелируют с изменениями в объеме производимой электроэнергии. Полученные результаты обеспечат основу для разработки моделей прогнозирования выработки и оптимизации режимов эксплуатации ГТУ.

1.1.2 Основные типы данных

При анализе данных в статистике принято выделять несколько ключевых типов информации, каждый из которых имеет свою специфику и требует определённого подхода при обработке. Количественные данные выражаются числами и отражают измеримые характеристики. Они могут быть непрерывными (например, рост или доход) или дискретными (например, количество детей в семье). С такими данными можно выполнять математические операции, рассчитывать средние значения, а также применять методы корреляции и регрессии.

К качественным данным относятся категориальные переменные, которые описывают нечисловые признаки, такие как уровень образования, семейное положение или место жительства. Эти данные обычно преобразуются в числовую форму с помощью кодирования или анализируются через таблицы распределения частот и сопряжённости.

Ещё один важный тип — временные ряды, представляющие собой данные, собранные в разные моменты времени. Например, можно отслеживать изменение эмоциональной удовлетворённости в браке на протяжении нескольких лет. Хотя подобная информация используется реже при прогнозировании таких событий, как развод, она позволяет выявлять тенденции и циклические изменения в семейных отношениях.

Правильное определение типа данных играет ключевую роль в выборе инструментов анализа, что в свою очередь обеспечивает достоверность выводов и их практическую значимость.

1.1.3 Ключевые задачи статистического анализа

Статистический анализ данных решает важные задачи, позволяющие выявлять закономерности, прогнозировать развитие событий и принимать обоснованные решения. В исследованиях, связанных с изучением вероятности развода, эти задачи становятся особенно значимыми, поскольку затрагивают сложные аспекты семейных отношений и дают возможность формировать эффективные подходы к их укреплению.

Прогностическая функция статистики направлена на построение моделей, способных предсказывать будущие исходы на основе текущих и исторических данных. В данном случае она используется для оценки риска расторжения брака с учетом таких факторов, как разница в возрасте между супругами, уровень дохода, эмоциональная вовлеченность. Для этих целей применяются регрессионные методы, алгоритмы машинного обучения и нейронные сети, что позволяет создавать надежные модели прогнозирования, полезные как для научных исследований, так и для практической работы специалистов.

Еще одной важной задачей является выявление взаимосвязей между различными переменными. Анализ зависимостей помогает определить, какие именно факторы наиболее сильно влияют на вероятность развода. Например, можно установить, что высокий уровень доверия и взаимопонимания в паре снижает риск распада семьи, тогда как наличие социальных или культурных различий, напротив, увеличивает его. Такие выводы позволяют глубже понять внутренние механизмы семейных конфликтов и разрабатывать рекомендации по повышению устойчивости брака.

Третьей ключевой задачей является оптимизация процессов принятия решений. Она заключается в использовании данных для повышения эффективности действий и планирования мер поддержки. Применительно к теме разводов это может означать создание профилактических программ,

ориентированных на группы повышенного риска. С помощью статистического анализа можно выделить такие группы и предложить им целевую помощь — от психологического сопровождения до образовательных курсов по управлению семейными отношениями. Это не только улучшает качество жизни семейных пар, но и способствует снижению общего уровня разводов в обществе.

Итак, статистический анализ выполняет три основные функции: прогнозирование, выявление причинно-следственных связей и оптимизацию действий. Все они играют важную роль при изучении устойчивости брака, обеспечивая как научную глубину понимания проблемы, так и ее практическое решение.

1.2 Основные методы статистической обработки данных

1.2.1 Описательная статистика

Описательная статистика — это базовый инструмент анализа данных, который позволяет систематизировать и представить информацию в наглядной форме. Этот метод помогает получить первичное представление о наборе данных, выявить их основные характеристики и подготовить почву для более глубокого исследования. В контексте изучения вероятности развода описательная статистика особенно важна, так как она позволяет проанализировать ключевые признаки, такие как возрастной разрыв между супругами, уровень дохода или степень приверженности к браку. Это помогает понять распределение этих факторов и их потенциальное влияние на вероятность расторжения брака.

Среднее значение отражает центральную тенденцию данных, то есть показывает типичное значение в выборке. Оно рассчитывается путем

суммирования всех значений и деления на их количество (1.1).

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (1.1)$$

где \bar{x} – среднее значение, x_i – индивидуальное значение признака (например, возрастной разрыв), n – количество наблюдений.

Медиана — это значение, которое делит упорядоченный набор данных на две равные части. Если данные расположены по возрастанию, медиана будет находиться точно посередине. Медиана менее подвержена влиянию выбросов, чем среднее значение, поэтому она особенно полезна при анализе данных с асимметричным распределением, таких как уровень дохода или количество детей от предыдущих браков. Например, медиана дохода может дать более точное представление о типичном уровне благосостояния, не искажаясь редкими случаями чрезвычайно высоких или низких доходов.

Мода определяет наиболее часто встречающееся значение в данных. Она особенно полезна для категориальных переменных, таких как уровень образования или религиозная принадлежность супругов. Например, мода в категории уровня образования может показать, какой уровень образования чаще всего встречается среди участников исследования. Это помогает выявить доминирующие категории, которые могут быть связаны с вероятностью развода.

Дисперсия измеряет степень разброса данных относительно их среднего значения. Она показывает, насколько сильно значения отличаются друг от друга. Чем выше дисперсия, тем больше вариабельность в данных. Дисперсия вычисляется следующим образом (1.2).

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}), \quad (1.2)$$

где s^2 - дисперсия, x_i - значение признака, \bar{x} - среднее, n - объём выборки.

Стандартное отклонение является квадратным корнем из дисперсии и выражается в тех же единицах, что и исходные данные. Это делает его более

интерпретируемым по сравнению с дисперсией. Стандартное отклонение также показывает, насколько сильно данные отклоняются от среднего значения.

Описательная статистика служит основой для анализа данных, предоставляя ключевые характеристики выборки в удобной и интерпретируемой форме. Она включает в себя такие показатели, как среднее значение, медиана, мода, дисперсия и стандартное отклонение — все они позволяют понять, как распределены данные, и выявить их основные закономерности.

1.2.2 Методы проверки гипотез

Методы проверки гипотез играют важную роль в статистическом анализе, позволяя оценивать достоверность различий между группами или взаимосвязей между переменными. Эти методы помогают исследователям принимать обоснованные решения на основе данных и избегать ложных выводов, вызванных случайными отклонениями. Особенно актуальны они при изучении сложных социальных явлений, таких как вероятность развода, где необходимо выявлять значимые факторы, влияющие на устойчивость брака.

Одним из часто используемых методов является t-тест, который позволяет сравнивать средние значения двух групп. Например, с его помощью можно определить, отличается ли уровень дохода между парами, которые сохранили брак, и теми, которые развелись. Однако для корректного применения этого теста необходимо соблюдение некоторых условий — например, нормальное распределение данных и схожие дисперсии в группах.

Когда требуется сравнить более чем две группы, используется ANOVA, или дисперсионный анализ. Этот метод позволяет выявлять различия между средними значениями нескольких групп одновременно. Например, можно проанализировать, как уровень образования супругов влияет на их

финансовую совместимость. Если ANOVA показывает наличие значимых различий, дополнительно применяются пост-тесты для уточнения, между какими именно группами наблюдаются эти различия.

Для анализа связей между категориальными переменными широко используется критерий хи-квадрат. Он позволяет оценить, существует ли зависимость между двумя качественными признаками, например, между уровнем образования и частотой разводов. Метод основан на сравнении фактических и ожидаемых частот в таблице сопряженности. Хотя он достаточно прост в использовании, требует достаточного количества наблюдений в каждой категории для получения надежных результатов.

Таким образом, методы проверки гипотез, такие как t-тест, ANOVA и хи-квадрат, являются важными инструментами статистического анализа. Они позволяют выявлять значимые различия и связи между переменными, что особенно важно при изучении факторов, влияющих на вероятность развода. Использование этих подходов обеспечивает научную обоснованность выводов и служит основой для дальнейшего моделирования и прогнозирования.

1.2.3 Корреляционный анализ и регрессионное моделирование

К Корреляционный анализ и регрессионное моделирование относятся к числу ключевых методов статистики, которые позволяют изучать взаимосвязи между переменными и строить модели для прогнозирования. Эти подходы широко используются в исследованиях, направленных на выявление факторов, влияющих на различные социальные явления, включая вероятность развода. Они дают возможность не только оценить степень взаимодействия между признаками, но и предсказывать исходы на основе имеющихся данных.

Корреляционный анализ используется для того, чтобы определить, насколько сильно и в каком направлении связаны две переменные. Например, он позволяет выяснить, увеличивается или уменьшается вероятность развода

с ростом уровня приверженности в браке. Связь между числовыми переменными обычно измеряется с помощью коэффициента корреляции Пирсона (1.3), который принимает значения от -1 до 1. Значение, близкое к 1, указывает на сильную положительную связь, к -1 — на сильную отрицательную, а значение около нуля говорит об отсутствии линейной зависимости. Для порядковых или нечисловых данных применяется ранговая корреляция Спирмена, которая оценивает связь на основе ранговых позиций значений.

Например, корреляционный анализ может показать, что между уровнем эмоциональной вовлеченности и стабильностью брака существует умеренная отрицательная связь, то есть более высокий уровень привязанности связан с меньшей вероятностью развода. Однако важно помнить, что корреляция сама по себе не означает причинно-следственной связи. То есть даже если между двумя переменными наблюдается сильная зависимость, это не гарантирует, что одна из них вызывает изменения другой. Такие выводы требуют дополнительного анализа, например, с использованием регрессионных моделей.

Таким образом, корреляционный анализ служит важным инструментом для выявления взаимосвязей между факторами, влияющими на вероятность развода. Он помогает формировать гипотезы и подготавливает основу для дальнейшего моделирования, особенно при построении регрессионных уравнений, которые позволяют уже не просто описывать, но и предсказывать развитие событий на основе наблюдаемых данных.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (1.3)$$

где r – коэффициент корреляции, x_i и y_i – значение двух переменных (например, уровень любви и вероятность развода), \bar{x} и \bar{y} – их средние значения, n – число наблюдений.

Регрессионное моделирование является одним из ключевых методов статистического анализа, который используется для построения зависимости

между целевой переменной и одним или несколькими предикторами. Оно позволяет не только прогнозировать значения зависимой переменной на основе известных значений независимых переменных, но и оценивать влияние каждого фактора, что делает его важным инструментом как в научных исследованиях, так и в прикладных задачах.

В случае, когда целевая переменная имеет бинарную природу — например, принимает значение 0 или 1 — применяется логистическая регрессия. Этот метод рассчитывает вероятность наступления определённого исхода, основываясь на значениях входных факторов. Логистическая регрессия широко используется в различных областях, таких как медицина, экономика, маркетинг и другие, где важно прогнозировать принадлежность к тому или иному классу.

Одним из главных преимуществ логистической регрессии является возможность интерпретации полученных результатов. Модель предоставляет коэффициенты для каждой независимой переменной, которые указывают на силу и направление её влияния на целевую переменную. Также вычисляются показатели значимости, позволяющие определить, какие факторы действительно оказывают существенное влияние на исход.

Таким образом, регрессионное моделирование, включая логистическую регрессию, играет важную роль в анализе данных. Оно служит мощным инструментом для прогнозирования и объяснения сложных процессов, обеспечивая как количественную оценку вероятностей, так и качественный анализ факторов, влияющих на исследуемое явление.

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}, \quad (1.4)$$

где $P(y = 1)$ – вероятность развода, β_0 – свободный член, β_i – коэффициенты, x_i – предикторы (например, возрастной разрыв, экономическое сходство), e – основание натурального логарифма.

Корреляционный анализ и регрессионное моделирование тесно связаны между собой и выполняют разные, но взаимодополняющие функции в

статистическом исследовании. С помощью корреляционного анализа изучается степень взаимосвязи между переменными, определяется сила и направление этой связи. Это позволяет предварительно выделить те факторы, которые могут оказывать влияние на целевую переменную.

После выявления значимых связей регрессионное моделирование используется для построения математической зависимости между переменными. Оно дает возможность не только описать обнаруженные взаимосвязи в виде уравнения, но и прогнозировать значения целевой переменной на основе известных значений входных факторов.

Такой последовательный подход — от анализа корреляций к построению регрессионной модели — обеспечивает более глубокое понимание структуры данных и позволяет формировать обоснованные гипотезы о характере взаимосвязей. Кроме того, результаты этих методов служат основой для применения более сложных аналитических инструментов, таких как алгоритмы машинного обучения и нейросетевые модели, которые требуют предварительного отбора информативных признаков и понимания их влияния на конечный результат.

1.2.4 Методы кластеризации и классификации

Методы кластеризации и классификации занимают важное место в анализе данных, поскольку позволяют либо объединять объекты в группы на основе их сходства, либо предсказывать принадлежность объектов к уже известным категориям. Эти подходы находят широкое применение при исследовании сложных социальных процессов, включая выявление закономерностей в поведении и принятии решений. Они являются неотъемлемой частью аналитических исследований, направленных на изучение различных явлений.

Кластеризация представляет собой метод группировки данных по

принципу внутреннего сходства без использования заранее заданных меток. Основная цель этого подхода — обнаружить естественную структуру в данных, разделив их на компактные и однородные группы. Например, с помощью алгоритма K-means можно разделить наблюдения на кластеры, основываясь на таких характеристиках, как уровень дохода, степень эмоциональной вовлеченности или продолжительность отношений. Это позволяет выделить типичные профили среди исследуемых объектов и глубже понять особенности каждой группы.

Однако стоит отметить, что кластеризация не предназначена для прогнозирования, она скорее помогает в описательном анализе и интерпретации данных. Качество полученного разбиения оценивается с использованием специальных метрик, которые определяют степень компактности внутри кластеров и различий между ними. Таким образом, кластерный анализ может служить мощным инструментом для предварительного изучения данных и формирования гипотез для дальнейшего исследования. Приведу пример с оценкой качества разделения на множества объектов на группы (1.5).

$$J = \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \mu_k\|^2, \quad (1.5)$$

где J — общая сумма квадратов внутри кластеров, K — число кластеров, C_k — множество объектов в кластере k , x_i — данные, μ_k — центроид кластера k .

Классификация — это метод машинного обучения, в котором используется заранее известная информация о принадлежности объектов к тем или иным классам. Основной задачей классификации является построение модели, способной предсказывать категорию для новых данных на основе уже размеченных примеров. В отличие от кластеризации, здесь каждому объекту в обучающей выборке присвоена метка, что позволяет обучать алгоритмы распознавать закономерности и делать точные прогнозы.

В рамках анализа сложных явлений, таких как социальные процессы, классификация может применяться для определения вероятности наступления того или иного исхода. Например, можно обучить модель различать два состояния: "наступление события" и "отсутствие события", основываясь на множестве факторов, характеризующих исследуемый объект.

Для решения задачи классификации используются различные подходы, включая логистическую регрессию, метод ближайших соседей (KNN), случайный лес, а также нейросетевые архитектуры, такие как многослойный перцептрон (MLP) или сверточные сети (CNN). Эти методы позволяют учитывать взаимодействие между множеством входных параметров, строить гибкие модели и достигать высокой точности предсказаний. Кроме того, они помогают выявлять наиболее информативные признаки, влияющие на конечный результат, что важно с точки зрения интерпретации.

Методы кластеризации и классификации тесно связаны между собой и могут использоваться совместно. Кластеризация часто применяется на начальных этапах анализа для выявления естественной группировки данных, изучения их структуры и формирования гипотез. После этого полученные группы могут быть использованы как дополнительные признаки в моделях классификации, либо классификация может проводиться на основе выделенных кластеров.

Такое сочетание подходов позволяет не только глубже понять данные, но и строить более эффективные предсказательные модели. В современном анализе сложных процессов эти методы играют ключевую роль, обеспечивая как качественную интерпретацию, так и надежное прогнозирование.

1.2.5 Методы нейросетевых моделей

Нейросетевые модели относятся к числу наиболее эффективных и гибких методов машинного обучения, которые активно применяются в задачах

прогнозирования и анализа сложных зависимостей. Эти модели имитируют принципы обработки информации человеческим мозгом, используя сети взаимосвязанных элементов — нейронов, что позволяет им выявлять глубокие закономерности даже в неочевидных данных. В современных исследованиях они находят применение в самых разных областях, включая анализ поведенческих и социальных процессов.

Одним из базовых типов нейросетей является многослойный персептрон (MLP), представляющий собой сеть прямого распространения сигнала, состоящую из входного, одного или нескольких скрытых и выходного слоёв. Такая архитектура способна улавливать нелинейные зависимости между входными признаками и целевой переменной, что делает её особенно полезной в задачах классификации и регрессии. В процессе обучения модель использует функции активации, такие как ReLU или сигмоида, чтобы вводить нелинейность в преобразования данных, а параметры сети корректируются с помощью алгоритмов оптимизации, например, стохастического градиентного спуска.

Еще одним мощным подходом являются сверточные нейронные сети (CNN), которые традиционно используются для анализа изображений, но также могут быть адаптированы под работу с табличными данными. Сверточные слои позволяют выделять локальные особенности и взаимосвязи между признаками, что особенно важно при работе с многомерными наборами данных. Например, CNN может эффективно учитывать комбинированное влияние различных факторов на конечный результат, обеспечивая более точное прогнозирование.

Для предотвращения переобучения, особенно при ограниченном объеме данных, в нейросетях применяются методы регуляризации, такие как dropout, batch normalization и L2-регуляризация. Эти механизмы обеспечивают устойчивость моделей и их способность обобщать на новых данных, что делает нейросети конкурентоспособными по сравнению с классическими статистическими методами.

Таким образом, использование нейросетевых моделей, таких как MLP и CNN, открывает широкие возможности для анализа сложных зависимостей в данных. Они позволяют строить высокоточные прогнозы, учитывая нелинейные и взаимосвязанные эффекты, что делает их важным инструментом в современных аналитических исследованиях.

1.2.6 Примеры практического применения статистических методов

Статистические методы играют важную роль в анализе технических процессов и широко используются для изучения факторов, влияющих на эффективность энергетического оборудования. В контексте исследования объема выработки электроэнергии газовой турбиной эти методы позволяют выявлять ключевые параметры, от которых зависит уровень генерации, а также строить модели для прогнозирования и оптимизации работы оборудования.

Одним из первых этапов анализа является применение описательной статистики, которая позволяет получить общее представление о характеристиках данных. Например, вычисление среднего значения и стандартного отклонения температуры на входе в турбину дает понимание типичных режимах работы и степени их вариабельности. Это помогает оценить, насколько стабильны условия эксплуатации и как они могут влиять на производительность установки.

Для выявления взаимосвязей между параметрами используется корреляционный анализ. Он может показать, например, что с увеличением давления на входе в турбину наблюдается рост объема выработки электроэнергии, тогда как при повышении температуры окружающей среды эффективность снижается. Такие выводы полезны для инженеров при планировании режимов работы и модернизации оборудования.

Методы проверки гипотез, такие как t-тест или ANOVA, применяются для сравнения эффективности турбины в различных условиях. Например,

можно проверить гипотезу о том, отличается ли средний объем выработки электроэнергии в разные сезоны года или при использовании разных видов топлива. Полученные результаты позволяют принимать обоснованные решения по оптимизации технологических процессов.

Регрессионное моделирование используется для построения математической зависимости объема выработки электроэнергии от ключевых факторов — таких как температура, давление, влажность воздуха и частота вращения вала. Линейная или нелинейная регрессия позволяет не только количественно оценить влияние каждого параметра, но и предсказывать выходные показатели на основе текущих значений входных переменных. Это особенно важно при управлении энергоблоками и планировании мощностей.

Наконец, для более точного прогнозирования и анализа сложных зависимостей применяются методы машинного обучения, включая случайный лес, градиентный бустинг и нейросетевые модели. Они способны учитывать нелинейные эффекты и взаимодействия между факторами, обеспечивая высокую точность предсказаний. Например, модель может быть обучена на исторических данных для прогноза суточного объема выработки электроэнергии на основе погодных условий, состояния оборудования и режимов эксплуатации.

Таким образом, статистические методы, начиная с описательного анализа и заканчивая сложными моделями машинного обучения, находят широкое применение при исследовании факторов, влияющих на объем выработки электроэнергии газовой турбиной. Эти подходы позволяют не только выявлять ключевые параметры, но и создавать эффективные инструменты прогнозирования и управления, что имеет важное практическое значение в энергетике.

1.3 Язык программирования R и его возможности для статистического анализа

1.3.1 Основные преимущества языка R

Язык программирования R является одним из ведущих инструментов для статистического анализа и обработки данных. Его популярность обусловлена широким функционалом, гибкостью и поддержкой со стороны научного и аналитического сообщества, что делает его особенно удобным для исследований, связанных с анализом сложных зависимостей.

Одним из ключевых достоинств R является наличие огромного количества пакетов, охватывающих практически все аспекты анализа данных. Например, такие библиотеки, как `tidyverse`, позволяют эффективно очищать, преобразовывать и анализировать данные, а `caret` и `mlr` предоставляют универсальные интерфейсы для построения моделей машинного обучения. Для работы с нейросетями доступны пакеты `keras` и `tensorflow`, которые позволяют реализовывать сложные модели, включая многослойный персептрон (MLP) и сверточные нейронные сети (CNN). Это дает возможность разрабатывать современные предсказательные модели без необходимости создания алгоритмов с нуля.

Еще одной сильной стороной языка R являются его мощные средства визуализации. Благодаря таким пакетам, как `ggplot2` и `plotly`, можно строить как классические графики (гистограммы, диаграммы рассеяния), так и интерактивные визуализации. В исследованиях, где необходимо наглядно представить распределение признаков, показать корреляции или проанализировать результаты классификации, эти возможности играют важную роль.

Кроме того, R — это свободно распространяемый язык с открытым

исходным кодом, что делает его доступным для студентов, ученых и специалистов любого уровня. Открытость платформы способствует активному развитию сообщества, которое регулярно выпускает новые пакеты, документацию и примеры использования, что особенно ценно при работе с различными наборами данных, включая файлы формата CSV.

R также предлагает встроенную поддержку статистических методов, что особенно важно для задач анализа данных. С его помощью можно легко вычислять меры описательной статистики, проводить проверку гипотез с использованием t-теста, ANOVA или критерия хи-квадрат, а также строить регрессионные модели. Все это позволяет быстро переходить от сбора данных к их полноценному анализу.

Важной чертой R является его ориентированность на воспроизводимость исследований. С помощью таких инструментов, как `rmarkdown`, пользователь может объединять код, текст и результаты в одном документе, создавая отчеты, которые полностью отражают процесс анализа. Это значительно повышает прозрачность и точность научной работы.

Таким образом, язык R представляет собой мощный и гибкий инструмент для обработки, анализа и визуализации данных. Его богатые функциональные возможности, широкое сообщество и фокус на статистике и воспроизводимости делают его отличным выбором для использования в курсовых работах и научных исследованиях, направленных на изучение сложных зависимостей и построение прогнозных моделей.

1.3.2 Обзор ключевых библиотек

Язык R предлагает широкий набор библиотек, которые делают его мощным инструментом для статистического анализа и обработки данных. Эти пакеты позволяют эффективно выполнять как предварительную подготовку данных, так и сложный анализ с последующей визуализацией и

моделированием.

Одной из ключевых групп библиотек является tidyverse — набор взаимосвязанных инструментов, ориентированных на работу с данными. Он предоставляет удобные функции для фильтрации, преобразования и анализа таблиц, что позволяет значительно упростить этап подготовки данных к анализу.

В состав tidyverse входит библиотека dplyr, которая содержит основные функции для манипуляций с данными. С её помощью можно выполнять такие операции, как выборка строк, добавление новых переменных, группировка и агрегирование данных. Это делает её незаменимым инструментом при работе с любыми наборами данных.

Еще одной важной библиотекой является ggplot2, предназначенной для построения графиков и диаграмм. Она позволяет создавать высококачественные визуализации, наглядно отражающие распределение данных, связи между переменными и другие закономерности. Благодаря своей гибкости и возможностям настройки она широко используется для исследовательского анализа.

Для построения моделей машинного обучения применяется библиотека caret. Она предоставляет универсальный интерфейс для обучения и оценки моделей классификации и регрессии, поддерживает множество алгоритмов и методов проверки качества моделей, таких как кросс-валидация и разбиение на обучающую и тестовую выборки.

Также стоит отметить библиотеку forecast, которая используется при работе с временными рядами. Она содержит средства для анализа динамики данных и построения прогнозов с использованием таких методов, как ARIMA и экспоненциальное сглаживание.

Все эти библиотеки предоставляют широкие возможности для анализа данных, начиная от простой обработки и заканчивая построением сложных моделей. Их использование делает язык R удобным и эффективным инструментом для решения задач в различных областях.

1.3.3 Примеры использования R в обработке данных

Язык R широко применяется для обработки данных в различных областях благодаря своим мощным инструментам и гибкости. В контексте исследования вероятности развода R демонстрирует свои возможности через конкретные примеры использования, которые иллюстрируют его практическую ценность для анализа данных, таких как датасет.

Одним из примеров является предварительная обработка данных с использованием библиотеки `dplyr`. Например, в работе данные из датасета очищаются от пропусков с помощью функции `mutate`, где пропущенные значения заменяются медианой, а выбросы удаляются с использованием метода IQR (межквартильный размах). Это позволяет подготовить данные для дальнейшего анализа, устраняя искажения, которые могли бы повлиять на результаты.

Еще одним примером служит визуализация данных с помощью `ggplot2`. В данном исследовании гистограммы и боксплоты строятся для анализа распределения таких признаков, как возрастной разрыв или экономическое сходство, с учетом целевой переменной — вероятности развода. Эти графики помогают исследователям визуально оценить закономерности и аномалии, что упрощает интерпретацию данных.

R также используется для построения предсказательных моделей с применением библиотеки `caret`. Например, логистическая регрессия применяется для прогнозирования вероятности развода на основе признаков, таких как уровень приверженности и социальные различия. Использование кросс-валидации в `caret` позволяет оценить точность модели, обеспечивая надежность предсказаний.

Кроме того, R применяется для реализации нейросетевых моделей через библиотеку `keras`. В работе многослойный персептрон (MLP) обучается на масштабированных данных для классификации пар по вероятности развода.

Это демонстрирует способность R работать с современными методами машинного обучения, что особенно ценно для обработки сложных многомерных данных.

Наконец, R используется для создания отчетов и воспроизводимости результатов с помощью `rmarkdown`. Все шаги анализа, включая загрузку данных, обработку и моделирование, могут быть задокументированы в одном документе, что упрощает проверку и повторение исследования, что важно для научной работы.

Таким образом, использование R в обработке данных, включая очистку, визуализацию, моделирование и документирование, делает его эффективным инструментом для анализа вероятности развода и подобных задач.

1.4 Программное обеспечение Glarus BI

1.4.1 Общая характеристика BI-системы

BI-система, или система бизнес-аналитики, представляет собой комплекс программных инструментов, предназначенный для обработки данных с целью их преобразования в полезную информацию для анализа и принятия решений. Такие системы позволяют собирать данные из различных источников, обрабатывать их, выявлять закономерности и представлять результаты в удобной для пользователя форме. Они находят применение во многих областях, где требуется системный подход к анализу информации.

Одной из ключевых особенностей BI-систем является способность интегрировать данные из разных источников, объединяя их в централизованное хранилище. Это позволяет работать с разнородными наборами данных, обеспечивая целостность информации. Также важна автоматизация процессов анализа, которая позволяет регулярно обновлять

отчеты, пересчитывать метрики и поддерживать актуальность данных без значительного участия аналитика.

Еще одной сильной стороной таких систем являются интерактивные дашборды, которые предоставляют пользователям возможность визуализировать данные, фильтровать их по интересующим параметрам и получать ответы на запросы в режиме реального времени. Это делает BI-инструменты удобными как для экспертов, так и для лиц, принимающих решения, особенно при необходимости быстро реагировать на изменения.

BI-системы поддерживают несколько типов аналитики: описательную, диагностическую, предсказательную и предписывающую. Описательная аналитика используется для обобщения имеющихся данных, диагностическая — для выявления причин наблюдаемых явлений, предсказательная — для моделирования будущих сценариев развития событий, а предписывающая — для рекомендаций по улучшению ситуации на основе анализа возможных действий.

Для реализации сложных статистических и машинных методов такие системы могут интегрироваться с языками программирования, такими как R или Python. Это расширяет их аналитические возможности и позволяет использовать в задачах, требующих глубокого моделирования и прогнозирования.

Кроме того, BI-системы отличаются высокой степенью масштабируемости и гибкости, что позволяет применять их как в небольших проектах, так и в крупных аналитических исследованиях. Их можно адаптировать под конкретные задачи и потребности пользователей, что делает их универсальным инструментом для работы с данными.

Таким образом, BI-системы играют важную роль в современном анализе данных, сочетая в себе возможности интеграции, автоматизации, визуализации и прогнозирования. Они обеспечивают эффективную поддержку принятия решений в самых разных областях, включая научные исследования и управление проектами.

1.4.2 Основные функциональные возможности

BI-системы обладают широким спектром функциональных возможностей, которые позволяют эффективно работать с данными на всех этапах анализа. Эти функции включают в себя импорт данных, их визуализацию и проведение различных видов аналитики, что делает BI-системы мощным инструментом для решения разнообразных задач.

Одной из ключевых возможностей является импорт данных, который позволяет собирать информацию из различных источников — от файлов формата CSV до корпоративных баз данных и облачных сервисов. Это обеспечивает гибкость при работе с разнородными наборами данных и возможность их объединения в единую систему. Например, можно загрузить данные из внешних файлов, объединить их с информацией из других источников и подготовить к дальнейшему анализу без потери целостности информации.

Визуализация данных — еще одна важная функция BI-систем. Она предоставляет пользователю средства для построения графиков, диаграмм и интерактивных дашбордов, что помогает быстро находить закономерности и выявлять аномалии. С помощью визуальных инструментов можно строить распределения значений, анализировать взаимосвязи между переменными и наблюдать за изменениями показателей в динамике. Интерактивные элементы, такие как фильтры и срезы, позволяют детализировать данные и сосредоточиться на наиболее значимых аспектах анализа.

Функции аналитики охватывают широкий диапазон методов — от простого суммирования данных до сложного прогнозирования. С помощью описательной аналитики можно рассчитывать основные статистические метрики, такие как средние значения или стандартные отклонения. Диагностическая аналитика позволяет выявлять причины наблюдаемых явлений, а предсказательная — строить модели, которые прогнозируют

развитие событий на основе исторических данных. Все эти подходы обеспечивают комплексный анализ и помогают принимать более обоснованные решения.

Кроме того, BI-системы могут интегрироваться с другими аналитическими инструментами, такими как R или Python, что позволяет использовать их для реализации сложных моделей машинного обучения и статистического анализа. Это делает их совместимыми с современными методами обработки данных и расширяет возможности как исследовательского, так и прикладного анализа.

Таким образом, функционал BI-систем охватывает все этапы работы с данными — от сбора и подготовки до визуализации и прогнозирования. Благодаря этому они находят применение в самых разных областях, обеспечивая поддержку принятия решений и углубленное изучение сложных процессов.

1.4.3 Примеры использования Glarus BI для анализа данных

В Glarus BI демонстрирует свои аналитические и визуализационные возможности на практике, применяясь в различных сферах для обработки данных, выявления закономерностей и поддержки принятия решений. Ниже приведены примеры использования системы в разных областях, которые иллюстрируют её функциональные возможности.

В сфере управления недвижимостью Glarus BI интегрировалась с CRM-системой и Google Sheets для автоматизации анализа финансовых показателей. Система собирала данные о доходах, расходах и загрузке объектов, после чего рассчитывала ключевые метрики и формировала отчеты. Это позволило оперативно отслеживать уровень заполняемости апартаментов, сезонность спроса и распределение прибыли по периодам, что упрощало управление и планирование ресурсов.

В медицинской сфере Glarus BI применялась для анализа информации из системы 1С Polyclinic. После внедрения система помогла выявить слабые места в процессах обслуживания пациентов, например, неравномерное распределение нагрузки между специалистами. Интерактивные дашборды позволили отслеживать поток клиентов, анализировать источники привлечения и корректировать график работы врачей. В результате удалось повысить эффективность работы клиник и оптимизировать предоставление услуг.

В маркетинговой аналитике Glarus BI интегрировалась с АМО CRM и рекламными кабинетами, такими как Facebook. Система выявила проблему "потерянных лидов" — обращений, которые не попадали в систему учета из-за ошибок в передаче данных. Для решения этой задачи была настроена автоматическая синхронизация и реализованы уведомления в Telegram, что значительно повысило качество работы с клиентами и увеличило конверсию.

Применительно к исследованию факторов, влияющих на вероятность развода, Glarus BI могла бы быть использована для построения интерактивного дашборда, который визуализирует связи между различными параметрами. Например, можно было бы представить графики, отражающие взаимосвязь между возрастным разрывом и частотой расторжения брака, или показать распределение уровня приверженности среди пар. Такие визуализации позволяют наглядно продемонстрировать зависимости, облегчить интерпретацию результатов и помочь в принятии обоснованных решений на основе данных.

1.5 Сравнение возможностей R и Glarus BI

1.5.1 Преимущества R

Язык R и BI-системы, такие как Glarus BI, оба являются полезными инструментами в анализе данных, однако у R есть ряд важных преимуществ, связанных с его гибкостью, аналитической мощностью и возможностями для построения сложных моделей.

Одним из ключевых достоинств R является его гибкость. В отличие от большинства BI-систем, где процессы зачастую стандартизированы и ориентированы на автоматизацию, R предоставляет пользователю полный контроль над каждым этапом анализа — от загрузки и очистки данных до построения модели и интерпретации результатов. Например, с помощью пакета `dplyr` можно выполнять точечные преобразования данных, фильтровать аномалии, создавать новые переменные и группировать данные с высокой степенью детализации. Кроме того, язык поддерживает написание пользовательских функций, что позволяет адаптировать вычисления под уникальные задачи исследования, чего сложно добиться в готовых аналитических платформах.

Еще одним важным преимуществом R является его мощность анализа, обусловленная глубокой интеграцией с классическими и современными статистическими методами. Язык изначально разрабатывался как инструмент для статистиков, поэтому он предлагает широкий набор встроенных функций и сторонних библиотек для анализа данных. Например, с помощью пакетов `stats` и `caret` можно легко проводить проверку гипотез (t-тест, ANOVA), строить регрессионные модели и оценивать корреляции между переменными. Это делает R особенно подходящим для исследований, требующих тонкого статистического анализа, в то время как в BI-системах такие операции часто ограничены базовым уровнем.

Кроме того, R демонстрирует превосходство в поддержке сложных моделей. С развитием машинного обучения и нейросетевых технологий R стал активно использоваться не только для традиционной статистики, но и для реализации продвинутых алгоритмов. Благодаря таким пакетам, как `keras`, `randomForest`, `xgboost` и другим, становится возможным обучение моделей, способных учитывать нелинейные зависимости и взаимодействия между признаками. Например, в научных работах R может применяться для построения многослойного персептрона или сверточных сетей, что открывает возможности для более точного прогнозирования и анализа сложных явлений. BI-системы же, хотя и позволяют визуализировать результаты и использовать некоторые преднастроенные модели, редко предоставляют возможность их глубокой настройки или самостоятельной реализации.

Таким образом, несмотря на удобство и наглядность BI-систем, язык R остается более мощным и гибким инструментом для проведения углубленного анализа данных, особенно когда речь идет о статистическом моделировании и применении современных методов машинного обучения.

1.5.2 Преимущества Glarus BI

Glarus BI, как современная система бизнес-аналитики, обладает рядом преимуществ перед языками программирования, такими как R, особенно в задачах, связанных с визуализацией данных и удобством использования. Эти особенности делают её более подходящим выбором в ситуациях, где важны скорость создания отчетов, наглядность и доступность анализа для пользователей без технической подготовки.

Одним из ключевых достоинств Glarus BI является возможность создания интерактивных отчетов. Система позволяет строить динамические дашборды, в которых пользователь может фильтровать данные, менять параметры и сразу видеть обновленные результаты. Например, можно создать

визуализацию, отображающую зависимость между различными факторами и частотой определённых исходов, при этом добавлять фильтры по другим переменным, чтобы детализировать анализ. Это делает информацию более понятной и полезной для специалистов, которым не нужно разбираться в коде или статистических моделях.

Еще одним важным преимуществом является простота использования. Glarus BI ориентирована на пользователей, не имеющих глубоких знаний в программировании или статистике. Её интерфейс интуитивно понятен — загрузка данных, построение графиков и формирование отчетов осуществляется с помощью простых действий, таких как перетаскивание полей и выбор типов диаграмм. Например, пользователь может быстро создать график, показывающий распределение определённых характеристик или взаимосвязь между переменными, без необходимости писать код. В отличие от этого, в R аналогичные действия требуют написания скриптов, что может быть трудозатратным и сложным для новичков.

Кроме того, Glarus BI обеспечивает высокую степень автоматизации. Система позволяет настраивать автоматическое обновление данных и генерацию регулярных отчетов, что значительно экономит время при работе с часто изменяющимися наборами данных. Например, при поступлении новых записей система может самостоятельно обновлять дашборды и отчеты, предоставляя актуальную информацию заинтересованным сторонам. В R подобная автоматизация возможна, но требует дополнительной настройки и программирования, что увеличивает временные и технические затраты.

Таким образом, хотя R превосходит по мощности анализа и гибкости моделирования, Glarus BI выигрывает в плане удобства, скорости работы и доступности. Она особенно эффективна в случаях, когда необходимо быстро представить данные в наглядной форме и сделать их доступными для широкого круга пользователей, включая руководителей, аналитиков и специалистов без технической подготовки.

1.5.3 Возможность совместного использования инструментов

Совместное использование языка R и BI-системы Glarus BI позволяет объединить их сильные стороны, создавая комплексный подход к анализу данных. Такой интеграция делает возможным как глубокую статистическую обработку, так и наглядное представление результатов, что повышает эффективность аналитического процесса.

R выступает основным инструментом для выполнения сложного анализа и построения моделей. В рамках исследования он может использоваться на этапе предварительной обработки данных — для очистки пропусков, удаления выбросов, кодирования категориальных переменных и масштабирования признаков. Кроме того, R позволяет строить сложные модели машинного обучения, включая нейронные сети, такие как многослойный персептрон (MLP), которые способны учитывать нелинейные зависимости между входными факторами. После завершения анализа результаты, например, предсказанные значения вероятности определённого исхода или коэффициенты корреляции между переменными, могут быть экспортированы в формате CSV или другом табличном виде.

Эти данные затем передаются в Glarus BI, где становится возможным их визуализация и интерактивная работа. С помощью этой системы можно создавать дашборды, отображающие ключевые метрики и закономерности, выявленные в ходе анализа. Например, можно построить графики, показывающие влияние различных факторов на конечный результат, добавить фильтры для детализации и сделать информацию доступной для пользователей без технической подготовки. Это особенно ценно, когда результаты анализа должны быть представлены руководству, специалистам или заинтересованным лицам, которым важна наглядность и простота восприятия информации.

Кроме того, Glarus BI предоставляет возможность быстрого создания

отчетов на основе уже обработанных данных. Например, если в R была проведена проверка гипотез, например, t-тест или ANOVA, результаты этих исследований можно передать в BI-систему и представить в виде графиков, таблиц или диаграмм, которые легко читаются и интерпретируются. Также можно организовать автоматическое обновление отчетов, что особенно удобно при работе с данными, которые регулярно пополняются.

Таким образом, комбинация R и Glarus BI позволяет построить эффективный рабочий процесс: R используется для углубленного анализа, моделирования и обработки данных, а Glarus BI — для визуализации, интерпретации и распространения результатов среди широкой аудитории. Это сочетание обеспечивает баланс между мощностью аналитики и удобством представления данных, что делает его ценным подходом во многих исследовательских и прикладных задачах.

2. ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ ИССЛЕДОВАНИЯ

Данный раздел направлен на применение теоретических знаний и инструментария, которые были описаны в теоретическом разделе.

Исследования будут производиться на заранее сгенерированном синтетическом наборе данных.

Будет использован функционал языка R и BI-система Glarus BI для визуализации результатов.

2.1 Описание используемых данных

Сгенерированный набор данных представляет с собой 150 строк, в которых есть несколько характеристик, от которых будет зависеть выработка энергии.

- AmbientTemperature_C - Температура окружающей среды (°C). Обычно более низкая температура на входе компрессора повышает плотность воздуха и, следовательно, массовый расход, что положительно сказывается на выработке.
- AmbientPressure_kPa - Атмосферное давление (кПа). Более высокое давление также положительно влияет на массовый расход воздуха.
- RelativeHumidity_percent - Относительная влажность (%). Высокая влажность может незначительно снижать выработку из-за вытеснения кислорода водяным паром.
- FuelFlowRate_kg_s - Расход топлива (кг/с). Основной управляющий параметр; больше топлива (до определенного предела) — больше выработки.

- CompressorInletTemp_C - Температура на входе в компрессор (°C). Может отличаться от AmbientTemperature_C из-за систем предварительного охлаждения/подогрева или рециркуляции.
- TurbineEfficiency_percent - КПД турбины (%). Также подвержен деградации.
- GasTurbineAge_years - Возраст газовой турбины (лет). Косвенно отражает общий износ.
- MaintenanceCycle - Категориальный признак, указывающий на стадию цикла технического обслуживания (например, "Early" - сразу после ТО, "Mid" - середина цикла, "Late" - приближается ТО). Это хорошо подойдет для ANOVA.
- PowerOutput_MW - Выработка электроэнергии (МВт)

AmbientTemperature_C	AmbientPressure_kPa	RelativeHumidity_percent	FuelFlowRate_kg_s	CompressorInletTemp_C	CompressorEfficiency_percent	TurbineEfficiency_percent	GasTurbineAge_years	MaintenanceCycle	PowerOutput_MW
16.2	102.54	23.6	25.54	17.0	91.89	92.04	5	Mid	158.06
33.5	99.2	57.2	21.17	32.3	86.32	87.0	18	Mid	116.48
27.0	98.72	57.8	18.48	27.1	88.22	90.07	10	Early	115.34
23.0	100.45	64.6	28.13	22.6	91.77	91.77	6	Early	174.07
9.7	102.93	70.8	12.22	12.0	90.64	93.64	1	Mid	86.16
9.7	99.21	88.3	19.85	10.2	88.39	94.78	5	Mid	125.51
6.7	101.36	56.1	10.23	7.4	86.97	93.53	9	Early	84.41
31.0	101.81	42.6	19.37	30.2	89.31	89.34	12	Late	105.8
23.0	99.19	75.7	11.13	22.1	85.88	88.15	14	Early	69.55
26.2	101.64	39.0	12.38	25.4	94.03	94.91	2	Early	95.51
5.6	99.84	50.7	12.35	6.4	83.78	87.37	17	Late	69.5
34.1	101.16	25.5	22.98	33.3	90.34	90.35	14	Late	130.63
30.0	101.17	21.8	24.92	30.6	88.42	90.1	13	Late	138.95
11.4	100.68	87.4	21.67	15.5	89.15	89.13	9	Late	116.71
10.5	98.45	78.5	29.24	12.2	86.68	91.32	15	Mid	174.65
10.5	102.18	68.7	17.5	9.9	88.1	91.69	15	Late	98.27
14.1	99.6	48.6	15.71	16.5	88.64	89.68	11	Early	102.34
20.7	98.93	32.1	27.37	19.9	88.85	88.02	16	Late	147.77
18.0	98.2	31.0	14.47	13.9	87.07	88.63	17	Late	79.48
13.7	100.95	37.5	29.26	11.7	91.03	93.9	4	Mid	181.45
23.4	101.39	58.4	10.24	19.6	91.15	95.59	1	Early	80.7
9.2	98.08	70.0	29.4	8.5	89.51	94.82	8	Mid	179.68
13.8	100.56	66.2	10.86	13.8	87.36	87.93	17	Late	63.64
16.0	99.13	39.6	27.82	19.3	88.53	92.6	3	Mid	167.4
18.7	101.23	86.8	20.55	19.3	83.98	84.93	20	Early	116.0
28.6	98.87	71.7	29.86	28.1	88.61	87.9	15	Early	174.93
11.0	101.45	58.8	11.48	12.6	87.76	88.95	15	Mid	76.43
20.4	99.93	62.8	21.08	16.0	87.52	92.09	14	Mid	123.08
22.8	102.68	49.4	29.39	23.2	87.83	88.62	14	Late	162.9
6.4	98.69	37.3	20.46	7.9	91.92	94.54	2	Early	140.71
23.2	99.71	44.9	22.59	20.3	87.29	89.78	13	Late	123.27

Рисунок 1 – Предложенный датасет.

2.1.1 Источники данных

Как упомянуто ранее, источник данных — синтетически сгенерированный набор данных, который был сгенерирован с помощью

Python кода. Все было учтено так, как выглядело бы в реальной ситуации. То есть проделанная работа будет актуальна и для реальных наборов данных.

```

1 import pandas as pd
2 import numpy as np
3
4 # Задаем количество строк
5 n_samples = 150
6
7 # Устанавливаем seed для воспроизводимости
8 np.random.seed(42)
9
10 # Генерация данных
11 data = {
12     'AmbientTemperature_C': np.random.uniform(5, 35, n_samples), # от 5 до 35 °C
13     'AmbientPressure_kPa': np.random.uniform(98, 103, n_samples), # от 98 до 103 кПа
14     'RelativeHumidity_percent': np.random.uniform(20, 90, n_samples), # от 20% до 90%
15     'FuelFlowRate_kg_s': np.random.uniform(10, 30, n_samples), # от 10 до 30 кг/с
16     'GasTurbineAge_years': np.random.randint(1, 21, n_samples) # от 1 до 20 лет
17 }
18
19 df = pd.DataFrame(data)
20
21 # CompressorInletTemp_C – может быть чуть выше AmbientTemperature_C из-за потерь или чуть ниже из-за охлаждения
22 df['CompressorInletTemp_C'] = df['AmbientTemperature_C'] + np.random.normal(0, 2, n_samples)
23 df['CompressorInletTemp_C'] = np.clip(df['CompressorInletTemp_C'], 0, 40) # Ограничим
24
25 # Эффективности компрессора и турбины зависят от возраста и случайных факторов
26 # Базовая эффективность + снижение с возрастом + шум
27 df['CompressorEfficiency_percent'] = 92 - (df['GasTurbineAge_years'] * 0.3) + np.random.normal(0, 1.5, n_samples)
28 df['CompressorEfficiency_percent'] = np.clip(df['CompressorEfficiency_percent'], 80, 95)
29
30 df['TurbineEfficiency_percent'] = 95 - (df['GasTurbineAge_years'] * 0.4) + np.random.normal(0, 1.5, n_samples)
31 df['TurbineEfficiency_percent'] = np.clip(df['TurbineEfficiency_percent'], 82, 98)
32
33 # MaintenanceCycle – категориальный
34 maintenance_options = ['Early', 'Mid', 'Late']
35 # Сделаем так, чтобы старые турбины чаще были в 'Late' цикле (условно)
36 probs = []
37 for age in df['GasTurbineAge_years']:
38     if age < 7:
39         probs.append([0.6, 0.3, 0.1]) # Early, Mid, Late
40     elif age < 14:
41         probs.append([0.3, 0.4, 0.3])
42     else:
43         probs.append([0.1, 0.3, 0.6])
44 df['MaintenanceCycle'] = [np.random.choice(maintenance_options, p=p) for p in probs]
45
46 base_power = df['FuelFlowRate_kg_s'] * 5.5 + 10
47
48 # Коррекции
49 temp_effect = (20 - df['CompressorInletTemp_C']) * 0.3 # Холоднее = лучше
50 pressure_effect = (df['AmbientPressure_kPa'] - 100) * 0.5 # Выше давление = лучше
51 humidity_effect = (50 - df['RelativeHumidity_percent']) * 0.05 # Суше = чуть лучше
52 comp_eff_effect = (df['CompressorEfficiency_percent'] - 88) * 0.8 # Выше КПД компр = лучше
53 turb_eff_effect = (df['TurbineEfficiency_percent'] - 90) * 1.0 # Выше КПД турб = лучше
54 age_effect = -df['GasTurbineAge_years'] * 0.1 # Старше = хуже
55
56 maintenance_modifier = pd.Series(index=df.index, dtype=float)
57 maintenance_modifier[df['MaintenanceCycle'] == 'Early'] = 5
58 maintenance_modifier[df['MaintenanceCycle'] == 'Mid'] = 0
59 maintenance_modifier[df['MaintenanceCycle'] == 'Late'] = -8
60
61 df['PowerOutput_MW'] = [(base_power +
62     temp_effect +
63     pressure_effect +
64     humidity_effect +
65     comp_eff_effect +
66     turb_eff_effect +
67     age_effect +
68     maintenance_modifier +
69     np.random.normal(0, 3, n_samples))] # Случайный шум
70
71 # Ограничим выработку разумными пределами
72 df['PowerOutput_MW'] = np.clip(df['PowerOutput_MW'], 30, 200) # от 30 до 200 МВт
73

```

Рисунок 2 – Код для синтетического набора данных.

2.1.2 Структура и характеристики данных

Все переменные, за исключением одной — являются числовыми и выражены в непрерывной шкале. В данном наборе данных присутствует один категориальный признак (Время с момента тех.обслуживания)

```
# A tibble: 6 × 10
  AmbientTemperature_C AmbientPressure_kPa RelativeHumidity_perc...1 FuelFlowRate_kg_s CompressorInletTemp_C
1             16.2             103.             23.6             25.5             17
2             33.5             99.2             57.2             21.2             32.3
3             27             98.7             57.8             18.5             27.1
4             23             100.             64.6             28.1             22.6
5             9.7             103.             70.8             12.2             12
6             9.7             99.2             88.3             19.8             10.2
# i abbreviated name: 1RelativeHumidity_percent
# i 5 more variables: CompressorEfficiency_percent <dbl>, TurbineEfficiency_percent <dbl>,
#   GasTurbineAge_years <dbl>, MaintenanceCycle <chr>, PowerOutput_MW <dbl>
> print("Структура данных:")
[1] "Структура данных:"
> print(str(df_expenses))
spec_tbl_ [150 × 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ AmbientTemperature_C      : num [1:150] 16.2 33.5 27 23 9.7 9.7 6.7 31 23 26.2 ...
 $ AmbientPressure_kPa       : num [1:150] 102.5 99.2 98.7 100.5 102.9 ...
 $ RelativeHumidity_percent  : num [1:150] 23.6 57.2 57.8 64.6 70.8 88.3 56.1 42.6 75.7 39 ...
 $ FuelFlowRate_kg_s         : num [1:150] 25.5 21.2 18.5 28.1 12.2 ...
 $ CompressorInletTemp_C     : num [1:150] 17 32.3 27.1 22.6 12 10.2 7.4 30.2 22.1 25.4 ...
 $ CompressorEfficiency_percent: num [1:150] 91.9 86.3 88.2 91.8 90.6 ...
 $ TurbineEfficiency_percent  : num [1:150] 92 87 90.1 91.8 93.6 ...
 $ GasTurbineAge_years        : num [1:150] 5 18 10 6 1 5 9 12 14 2 ...
 $ MaintenanceCycle          : chr [1:150] "Mid" "Mid" "Early" "Early" ...
 $ PowerOutput_MW            : num [1:150] 158.1 116.5 115.3 174.1 86.2 ...
- attr(*, "spec")=
.. cols(
..   AmbientTemperature_C = col_double(),
..   AmbientPressure_kPa = col_double(),
```

Рисунок 3 – Отображение датасета

Из рисунка 3 можно сделать вывод, что датасет корректно загружен и с ним можно продолжать работу.

2.1.3 Предварительная обработка данных

Важно сделать предварительную обработку данных датасета. Процесс представляет с собой подготовку данных к будущему анализу. Выбросы, пропуски и некорректные значения могут поломать код программы, а в худшем случае — испортить результаты исследования без информирования ошибки компилятором.

Важным шагом для работоспособности алгоритма k-means является масштабирование данных. Используем язык R, чтобы привести данные к нужному виду, так как Евклидово расстояние требует приближенный диапазон значений для каждой характеристики.

AmbientTemperature_C	AmbientPressure_kPa	RelativeHumidity_percent	FuelFlowRate_kg_s	CompressorInletTemp_C
Min. :-1.57206	Min. :-1.7558	Min. :-1.60849	Min. :-1.69010	Min. :-2.0844
1st Qu.: -0.86690	1st Qu.: -0.9270	1st Qu.: -0.82387	1st Qu.: -0.86682	1st Qu.: -0.8082
Median :-0.08308	Median : 0.1317	Median : 0.02712	Median : 0.02057	Median :-0.1245
Mean : 0.00000	Mean : 0.0000	Mean : 0.00000	Mean : 0.00000	Mean : 0.0000
3rd Qu.: 0.93673	3rd Qu.: 0.8250	3rd Qu.: 0.85203	3rd Qu.: 0.94332	3rd Qu.: 0.9081
Max. : 1.73178	Max. : 1.6212	Max. : 1.63902	Max. : 1.55862	Max. : 1.8687
CompressorEfficiency_percent	TurbineEfficiency_percent	GasTurbineAge_years		
Min. :-2.31925	Min. :-2.71025	Min. :-1.60716		
1st Qu.: -0.74093	1st Qu.: -0.75548	1st Qu.: -0.90991		
Median :-0.03218	Median : 0.02897	Median :-0.03835		
Mean : 0.00000	Mean : 0.00000	Mean : 0.00000		
3rd Qu.: 0.79303	3rd Qu.: 0.78899	3rd Qu.: 0.83322		
Max. : 2.29482	Max. : 2.16642	Max. : 1.70478		

Рисунок 4 – Стандартизация данных

Заметим, что средние значение у всех столбцов равно 0, что является подтверждением стандартизации данных. Стандартизированный датасет (scaled_df) готов для дальнейшего анализа данных.

2.2 Исследовательский анализ данных (EDA – Exploratory Data Analysis)

Не менее важным этапом является исследовательский анализ данных (EDA). Он помогает получить основную структуру данных и выявить некоторые закономерности, также помогает проверить предположения о данных и выбрать подходящий метод для анализа.

2.2.1 Визуализация распределений

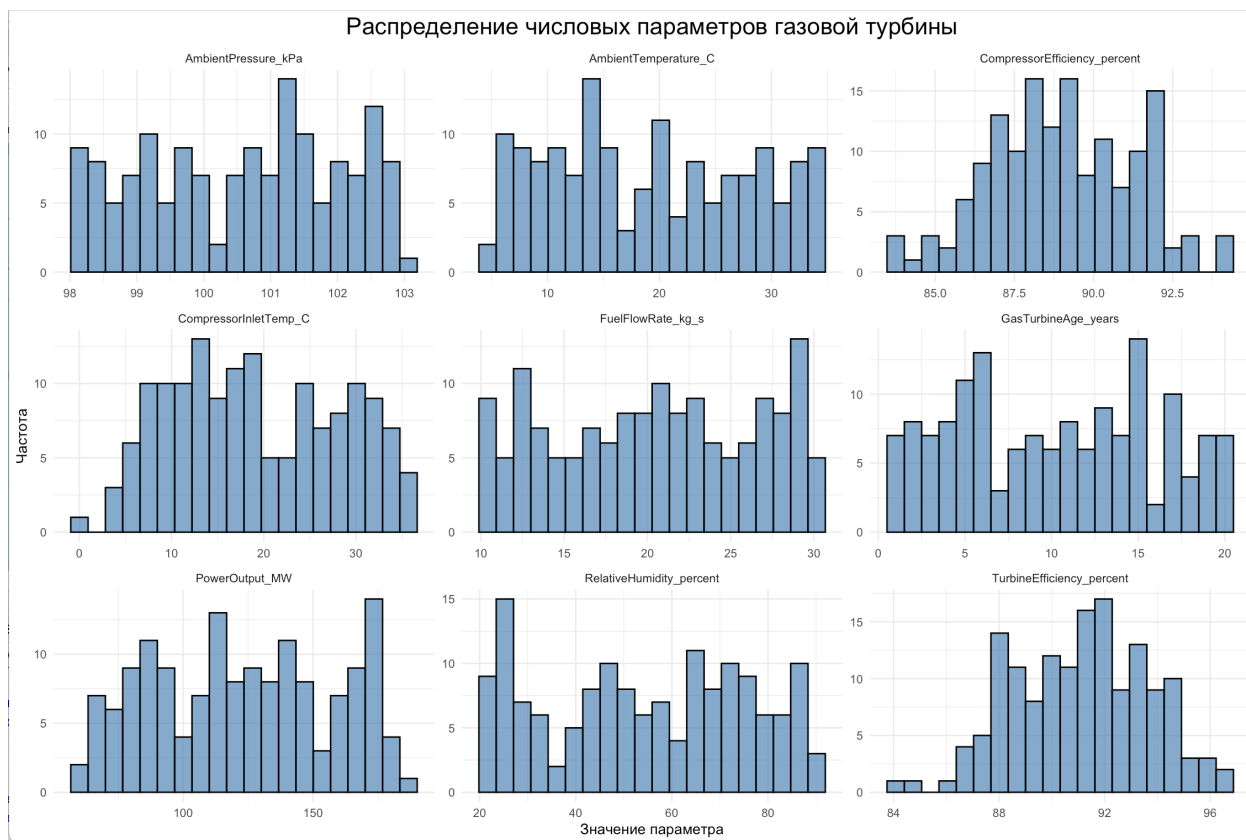


Рисунок 5 – Гистограммы столбцов

Потенциал для кластеризации: Мультимодальность, особенно в `PowerOutput_MW` и `GasTurbineAge_years`, а также разнообразие других распределений, говорит о том, что в данных, вероятно, существуют естественные группы (кластеры).

Факторы влияния: Визуально уже можно предположить, что разные комбинации этих параметров могут приводить к разным уровням выработки. Например, старые турбины с низким КПД, вероятно, будут иметь более низкую выработку.

Далее построим box plot для выявления потенциальных выбросов, будем производить на категориальном столбце

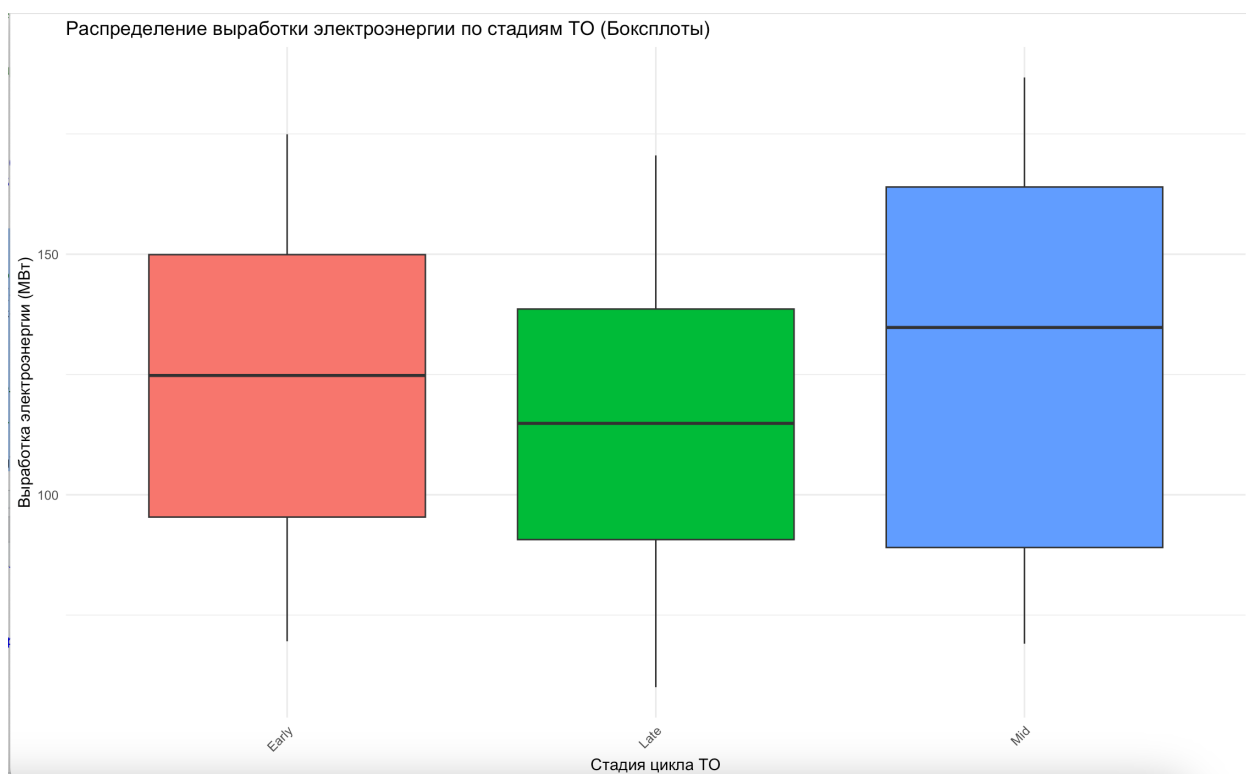


Рисунок 6 – Боксплот

Стадия Late: Значительный разброс. Сам бокс (межквартильный размах, IQR, содержащий средние 50% данных) достаточно широк, и "усы" также простираются на значительное расстояние. Это говорит о том, что хотя медиана и низкая, все еще существует большой разброс в производительности турбин на этой стадии. Некоторые все еще могут выдавать приличную мощность, но многие показывают низкую.

Стадия Early: Большой разброс. IQR (высота бокса) и общий размах "усов" также велики, возможно, даже чуть больше, чем у "Late" в терминах IQR. Это означает, что турбины сразу после ТО показывают в среднем хорошую выработку, но эта выработка также может сильно варьироваться.

Стадия Mid: Наибольший разброс (IQR и общий). Синий бокс самый "высокий", что указывает на наибольший межквартильный размах. "Усы" также простираются очень широко, особенно верхний. Это означает, что на средней стадии цикла ТО наблюдается самый большой разброс в выработке электроэнергии – есть как очень производительные экземпляры, так и те, чья производительность уже не так высока.

2.2.2 Корреляционный анализ

Корреляционный анализ поможет нам определить зависимости характеристик

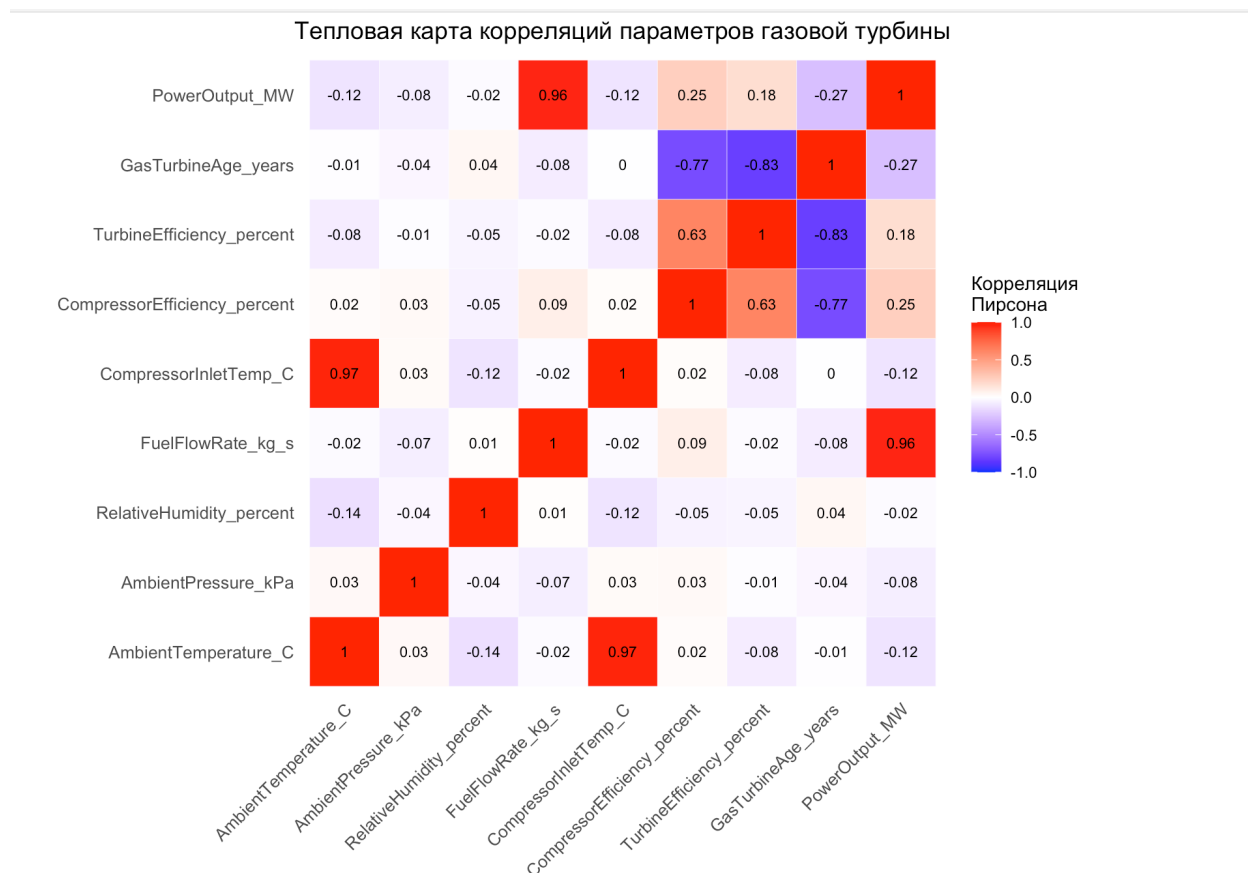


Рисунок 7 – Корреляционный анализ

Сделаем тепловую карту, которая отразит зависимости переменных между собой. По рисунку 7 можно сделать следующие выводы:

PowerOutput_MW и FuelFlowRate_kg_s (0.96): Это самая сильная и ожидаемая корреляция. Чем больше расход топлива, тем выше выработка электроэнергии. Это ключевой фактор управления мощностью.

CompressorInletTemp_C и AmbientTemperature_C (0.97): Очень сильная положительная корреляция. Температура на входе в компрессор практически полностью определяется температурой окружающей среды. Это логично, если нет значительных систем предварительного охлаждения/подогрева на входе.

TurbineEfficiency_percent и CompressorEfficiency_percent (0.63):

Умеренная положительная корреляция. КПД турбины и КПД компрессора имеют тенденцию изменяться в одном направлении. Это может быть связано с общим состоянием турбины: если один компонент в хорошем состоянии, то и другой, вероятно, тоже. Или же их деградация происходит параллельно.

`TurbineEfficiency_percent` и `GasTurbineAge_years` (-0.83): Сильная отрицательная корреляция. С увеличением возраста турбины ее КПД значительно снижается. Это отражает износ.

`CompressorEfficiency_percent` и `GasTurbineAge_years` (-0.77): Также сильная отрицательная корреляция. КПД компрессора также сильно падает с возрастом турбины.

`PowerOutput_MW` и `GasTurbineAge_years` (-0.27): Умеренная отрицательная корреляция. С возрастом турбины выработка электроэнергии имеет тенденцию снижаться, но эта связь не такая сильная, как между возрастом и КПД. Это говорит о том, что другие факторы (например, расход топлива) могут компенсировать падение КПД.

2.2.3 Выявление выбросов и трендов

Так как датасет является синтетически созданным, то выбросов быть не может, а значит данный пункт не обязателен для выполнения последующего анализа

2.3 Применение методов статистического анализа

Данный подраздел посвящен использованию классических статистических методов в качестве важного элемента общего процесса анализа данных. Они служат для более глубокого понимания структуры данных и помогают обосновать выводы, полученные с помощью методов машинного

обучения.

Статистические подходы, включая описательную статистику и проверку гипотез, выполняют вспомогательную, но значимую роль на различных этапах исследования. С их помощью можно изучить распределение признаков, выявить закономерности и оценить статистическую значимость наблюдаемых различий. Кроме того, такие методы способствуют интерпретации и проверке результатов, полученных при кластеризации данных, обеспечивая их дополнительное обоснование и повышая надежность анализа.

2.3.1 Описательная статистика

Проведем самую стандартную описательную статистику с помощью функции `summary()`

```
AmbientTemperature_C AmbientPressure_kPa RelativeHumidity_percent FuelFlowRate_kg_s CompressorInletTemp_C CompressorEfficiency_percent
Min. : 5.20      Min. : 98.03      Min. :20.80      Min. :10.23      Min. : 0.40      Min. :83.78
1st Qu.:11.47    1st Qu.: 99.24    1st Qu.:37.35    1st Qu.:15.24    1st Qu.:11.93    1st Qu.:87.34
Median :18.45    Median :100.78    Median :55.30    Median :20.64    Median :18.10    Median :88.94
Mean :19.19      Mean :100.59      Mean :54.73      Mean :20.51      Mean :19.22      Mean :89.01
3rd Qu.:27.52    3rd Qu.:101.79    3rd Qu.:72.70    3rd Qu.:26.25    3rd Qu.:27.43    3rd Qu.:90.80
Max. :34.60      Max. :102.95      Max. :89.30      Max. :29.99      Max. :36.10      Max. :94.18

TurbineEfficiency_percent GasTurbineAge_years PowerOutput_MW
Min. :84.16      Min. : 1.00      Min. : 60.01
1st Qu.:89.16    1st Qu.: 5.00    1st Qu.: 93.37
Median :91.17     Median :10.00     Median :123.17
Mean :91.10      Mean :10.22      Mean :123.09
3rd Qu.:93.11    3rd Qu.:15.00    3rd Qu.:150.30
Max. :96.64      Max. :20.00      Max. :186.73
```

Рисунок 8 – Описательная статистика

На рисунке 8 можно заметить основные статистические данные, такие как: Максимальное и минимальное значение, среднее значение, медиана, 1 и 3 (25%, 75% соответственно) квантили.

2.3.2 Проверка гипотез

2.3.3 Регрессионный анализ

Работа не подразумевает использование прогнозирования, а также требует “учителя”, что не соответствует фундаментальной логике работы поэтому пункт пропускается.

2.4 Машинное обучение в анализе данных

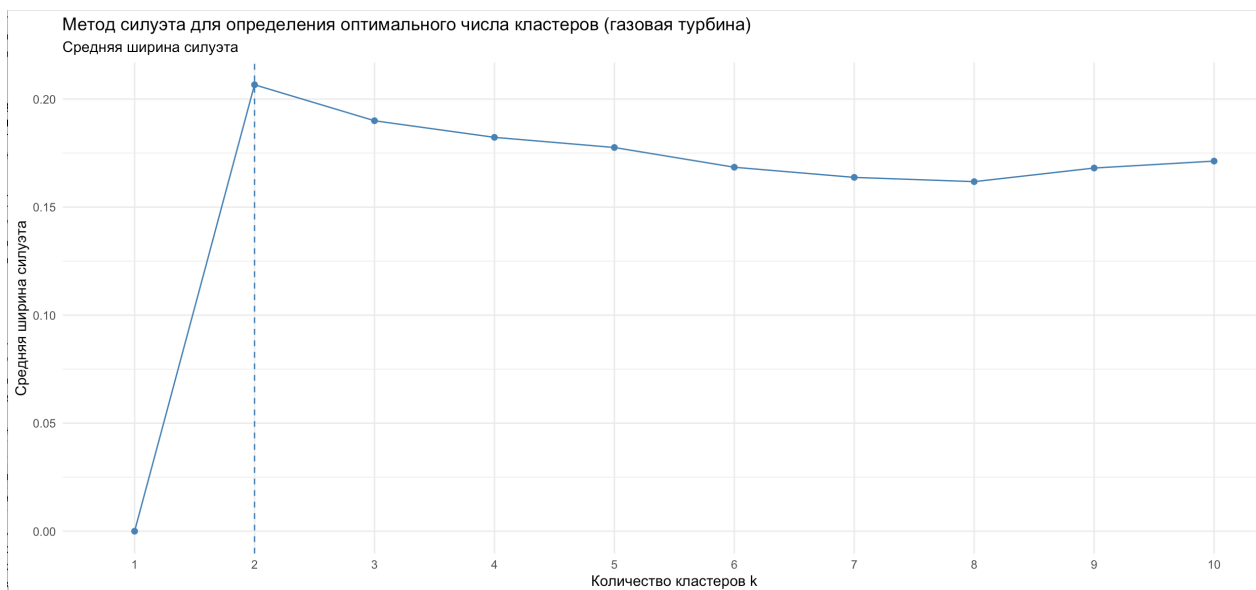
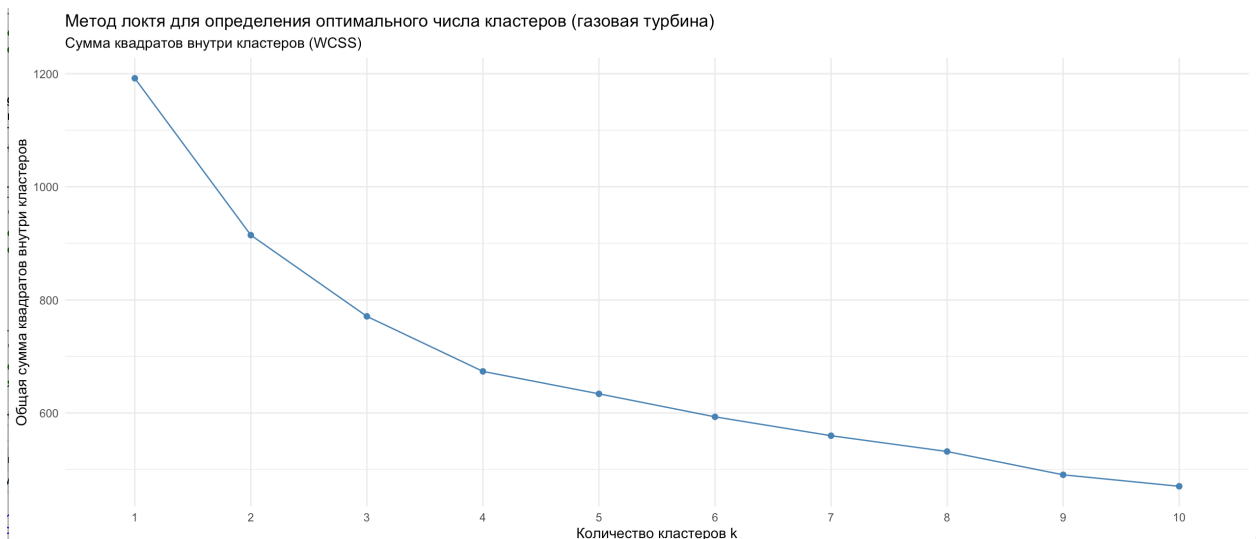
Является опциональным разделом для данной темы курсовой работы, в данном разделе понадобится только Кластеризация данных.

2.4.1 Классификация данных

Классификация данных по определению является методом обучения с “учителем”, целью которой является классификация объекта к уже определенному набору данных (“Учителю”). У нас стоит совершенно другой метод в рамках данной работы.

2.4.2 Кластеризация

В коде программы добавим графики метода локтя и силуэтов.



K-means clustering with 3 clusters of sizes 44, 41, 65

Cluster means:

	AmbientTemperature_C	AmbientPressure_kPa	RelativeHumidity_percent	FuelFlowRate_kg_s	CompressorInletTemp_C
1	1.0654465	0.10974422	-0.26541641	-0.14959035	1.0321539
2	-0.8876648	0.03999879	0.03498081	0.21591912	-0.8633718
3	-0.1613137	-0.09951840	0.15760168	-0.03493398	-0.1541005

	CompressorEfficiency_percent	TurbineEfficiency_percent	GasTurbineAge_years
1	0.5432668	0.4652853	-0.6167506
2	0.6538153	0.8375518	-0.7738640
3	-0.7801564	-0.8432643	0.9056223

Clustering vector:

```
[1] 2 3 1 1 2 2 2 1 3 1 3 1 1 3 3 3 3 3 2 1 2 3 2 3 3 3 3 2 3 3 3 1 1 3 3 2 3 3 3 2 2 3 3 1 2 1 3 2 1 3 1 3 1 1 2 2 3 3 3 2 3 2 3
[66] 1 2 3 2 1 1 2 2 1 3 1 1 2 2 2 1 1 3 2 3 3 1 3 1 2 2 1 1 3 1 2 3 2 3 2 3 1 3 3 1 3 3 1 2 2 2 3 3 1 3 3 1 3 1 1 3 2 2 3 1 1 2 3
[131] 3 2 2 3 1 3 1 1 3 1 1 2 3 3 3 3 1 2 3 2
```

Within cluster sum of squares by cluster:

```
[1] 215.4177 196.1951 359.3002
(between_SS / total_SS = 35.3 %)
```

Available components:

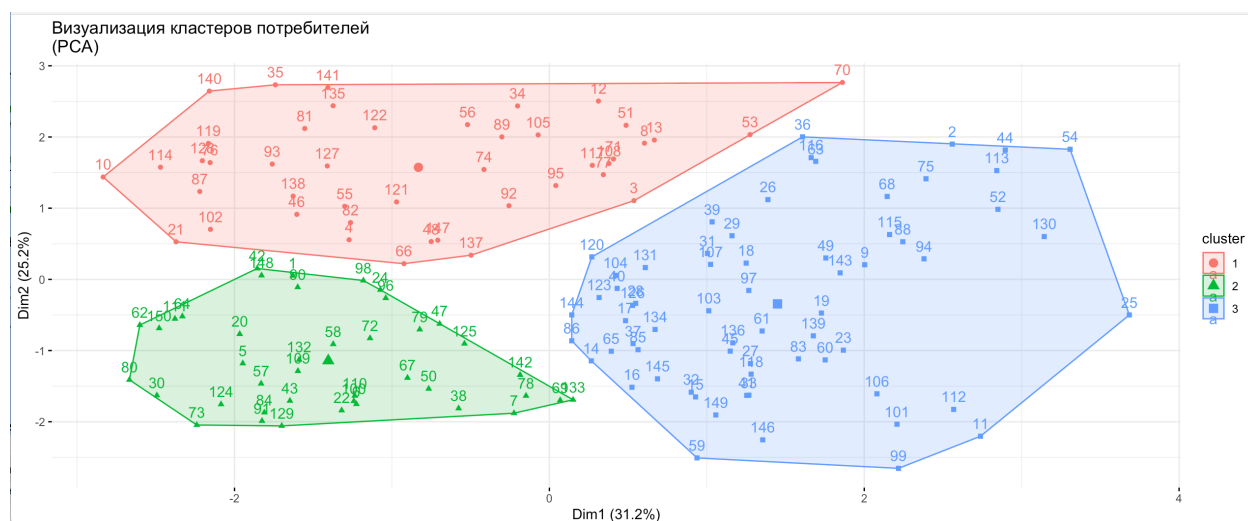
```
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

2.4.3 Прогнозирование временных рядов

Датасет не предоставляет данные о временных значениях, а значит данный пункт является технически невозможным, а значит не является обязательным для выполнения последующего анализа.

2.5 Визуализация данных

2.5.1 Визуализация в R



2.5.2 Интерактивные дашборды в Glarus BI

2.5.3 Сравнение методов визуализации

3. АВТОМАТИЗАЦИЯ И ОТЧЁТНОСТЬ В

АНАЛИЗЕ ДАННЫХ

3.1 Генерация отчётов в R

3.1.1 Обоснование необходимости автоматизации отчётов

3.1.2 Использование RMarkdown для создания отчетов

3.1.3 Экспорт отчетов в PDF, HTML, Word

3.2 Формирование интерактивных отчётов в Glarus BI

3.2.1 Различие между статичными и интерактивными отчетами

3.2.2 Создание дашбордов в Glarus BI

3.2.3 Экспорт отчетов в Glarus BI

3.3 Сравнение инструментов R и Glarus BI

3.3.1 Анализ сильных и слабых сторон инструментов

3.3.2 Возможности интеграции R и Glarus BI

3.3.3 Применимость инструментов для различных типов задач

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

- 1.1. Д.Ю. Кузнецов, Т.Л. Трошина Кластерный анализ и его применение – Ярославский педагогический вестник, 2006. – 1 с.
- 1.2. (Епифанцева, М. Я. Вероятностно-статистическое обеспечение обработки информации: учебно-методическое пособие / М. Я. Епифанцева. — Омск: ОмГУПС, 2023 — Часть 1: Корреляционный анализ — 2023. — 28 с. — Текст: электронный // Лань: электронно-библиотечная система. — URL: <https://e.lanbook.com/book/419273> (дата обращения: 08.05.2025). — Режим доступа: для авториз. пользователей. — С. 6.).
- 1.3. А.П. Баврина, И.А.Борисов Современные правила применения корреляционного анализа – ФГБОУ ВО «Приволжский исследовательский медицинский университет» Минздрава России, Нижний Новгород, 27.01.2021 – 2 с.
- 1.4. Яковлева Н.А., доцент Черникова В. С. Орловский Государственный Аграрный Университет Россия, г. Орел Корреляционный анализ уровня оплаты труда 2014 – 1 с.
- 1.5. Корреляция, корреляционная зависимость. Математическая статистика для психологов. — URL: <https://statpsy.ru/correlation/correlation/> (дата обращения: 08.05.2025).
- 1.6. Кластерный анализ: Базовые концепции и алгоритмы Долгодворова Е.В. 2018 – 2 с.
- 1.7. Иерархическая кластеризация — Викиконспекты <https://neerc.ifmo.ru/wiki/index.php> (дата обращения: 08.05.2025).

- 1.8. Прикладная статистика. Основы эконометрики: В 2 т. 2-е изд., 52 испр. - Т.2: Айвазян С.А. Основы эконометрики. - М.: ЮНИТИ-ДАНА, 2001. - 432 с.
- 1.9. Кластеризация — Викиконспекты (дата обращения: 08.05.2025).
<https://neerc.ifmo.ru/wiki/index.php>
- 1.10. Документация язык R. [Электронный ресурс]. <https://www.r-project.org/about.html> (дата обращения: 08.05.2025).
- 1.11. Язык программирования R для статистической обработки данных Гибадуллина Дарья Анатольевна/ Gibadullina Daria Anatolievna - Уральский филиал Финансового университета 17.31.2024
- 1.12. С.С. Задорожный Статистическая обработка данных на языке R. Учебно-методическое пособие.
<https://cmp.phys.msu.su/sites/default/fileshttps://cmp.phys.msu.su/sites/default/files>
- 1.13. Обзор документации. Документация Glarus BI [Электронный ресурс]. <https://glarus-bi.ru/docs/> (дата обращения: 02.05.2025).
- 1.14. Визуализация. Документация Glarus BI [Электронный ресурс]. <https://glarus-bi.ru/docs/> (дата обращения: 01.05.2025).
- 1.15. Дашборды. Документация Glarus BI [Электронный ресурс]. <https://glarus-bi.ru/docs/> (дата обращения: 01.05.2025).

ПРАКТИЧЕСКАЯ ЧАСТЬ

- 2.1 Marriage and Divorce Dataset Kaggle [Электронный ресурс].
«Marriage and Divorce Dataset» (дата обращения: 15.04.2025).
- 2.2 Онлайн-документация Glarus BI [Электронный ресурс].
<https://glarus-bi.ru/docs/> (дата обращения: 01.05.2025)
- 2.3 Онлайн-документация R [Электронный ресурс]. <https://www.r-project.org/about.html> (дата обращения: 01.05.2025)

ПРИЛОЖЕНИЕ А