ВВЕДЕНИЕ

Современный рынок лодок и катеров представляет собой динамичную и конкурентную отрасль, где успешность бизнеса зависит от множества факторов, таких как размеры судов, характеристики, их популярность и привлекательность для потенциальных покупателей.

Анализ этой рыночной сферы становится ключевым инструментом для компаний, занимающихся производством и продажей водных транспортных средств.

Цель данного исследования — применение методов анализа и прогнозирования для создания модели, способной разделить лодки на кластеры на основе их характеристик и размеров, для выявления групп схожих судов. При этом, мы также уделим внимание качественному оформлению документации и отчета, что является важным элементом профессиональной работы в сфере анализа данных.

В рамках данной курсовой работы будут решены следующие задачи:

* изучить научную и методическую литературу, связанную с анализом рынка водных средств транспорта, а также методами кластерного анализа;
* собрать необходимые данные о лодках и катерах, включая их размеры, характеристики и данные о просмотрах объявлений;
* применить язык программирования R для статистической обработки данных и работы с графикой для проведения кластеризации лодок на основе их размеров и других характеристик;
* создать модели прогнозирования разделения лодок на кластеры и качественный анализ результатов этой модели.

1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Предметная область исследования

Актуальность изучения рынка водного транспорта также связана с его долгосрочной устойчивостью и потенциалом для инноваций. С учетом изменения климата и усиливающегося экологического сознания, развитие экологически чистых технологий и устойчивых практик в судоходстве становится необходимостью. Это ведет к новым требованиям к кораблям и судовладельческим компаниям, что влияет на динамику рынка и создает потребность в анализе текущих тенденций.

Планета Земля — это планета большую часть, которой занимает Мировой океан, а именно 70,8 % ее территории. Так что по существу Земля — это водная планета.

В наше время водный транспорт распространен во всем мире. Большое количество грузоперевозок осуществляется по водным просторам. Так же свою роль играет пассажирское сообщение.

Россия же занимает большое пространство, требующее наличия достаточного количества транспортных коммуникаций.

По мере развития рыночной экономики в стране повышение эффективности транспортного процесса требует новых подходов к организации перевозок. Это привело к появлению нового направления — транспортной логистики.

Водный транспорт как составная часть более крупной системы, т.е. логистической цепи, привел к необходимости рассматривать его в разных аспектах. Транспортный сервис в современных условиях включает в себя не только собственно перегрузку грузов от поставщика потребителю, но и большое количество экспедиторских, информационных и трансакционных операций, услуг по грузопереработке, страхованию, охране и т.д. Такой подход способствует оптимальному выбору транспортных услуг, ибо качество перевозок, как правило, в большей мере отражается на общих расходах, чем себестоимость перевозок.

Водный транспорт принято разделять на глубоководное (океанское, морское) и внутреннее (речное).

Главное преимущество водного транспорта — это способность перевозить очень крупные грузы. При этом используется два типа судов: глубоководные (нуждаются в портах с глубоководными акваториями) и дизельные баржи (обладают большей гибкостью). Главными недостатками водного транспорта являются ограниченные функциональные возможности и небольшая его скорость.

Причина в том, что для доставки грузов в порты и из портов приходится использовать железные дороги или грузовики, за исключением случаев, когда и пункт отправления, и пункт назначения расположены на одном и том же водном пути. Водный транспорт, таким образом, отличающийся большой грузоподъемностью и незначительными переменными издержками, выгоден тем грузоотправителям, для которых важны низкие транспортные тарифы, а скорость доставки имеет второстепенное значение.

Типичными грузами для перевозки по внутренним водным путям являются руда, минеральное сырье, цемент, зерно и некоторые другие сельскохозяйственные продукты.

Возможности транспорта ограничены не только его привязкой к судоходным рекам и каналам, но и зависимостью от мощностей для погрузки-разгрузки и хранения таких насыпных грузов, а также растущей конкуренцией со стороны железных дорог, обслуживающих параллельные дороги. Мы узнали какие бывают виды судов, теперь можно перейти к внутреннему водному транспорту.

Россия обладает большой и разветвленной сетью речных путей и озер. Однако существенную роль она играет либо в тех регионах, где направления основных транспортно-экономических связей и речных путей совпадают (Волжско-Камский речной бассейн в европейской части России), либо в слабоосвоенных регионах с практически полным отсутствием альтернативных видов транспорта (Север и Северо-Восток страны).

В России насчитывается более 100 тысяч рек, имеющих общую протяжённость около 2,5 миллионов километров, из них свыше 500 тысяч километров пригодны для судоходства.

Различаются магистральные речные пути, обслуживающие международные связи, межрайонные, обеспечивающие перевозки грузов и людей между крупными районами внутри страны, и местные, обеспечивающие внутрирайонные связи.

Протяженность эксплуатируемых внутренних водных путей в России в последние десятилетия сокращается и в настоящее время составляет 89 тысяч километров, также на речном транспорте постоянно сокращается средняя дальность перевозки 1 т груза и в настоящее время с учетом всех видов речных сообщений она составляет менее 200 км.

Внутренние судоходные водные пути относятся к разным речным бассейнам. Преобладающую часть грузовых перевозок и грузооборота выполняют пароходства трех воднотранспортных бассейнов: Волжско-Камского, Западно-Сибирского и Северо-Западного.

Большая часть оборота речного транспорта приходится на европейскую часть страны. Важнейшей транспортной речной магистралью здесь является Волги со своим притоком Камой. На севере европейской части России заметную роль играют Северная Двина, Онежское и Ладожское озера, реки Свирь и Нева.

Большое значение для развития речного транспорта в стране имело создание единой глубоководной системы и строительство Беломоро-Балтийского, Волго-Балтийского, Московско-Волжского и Волго-Донского каналов.

В связи с освоением природных ресурсов на востоке страны возрастает транспортное значение Оби, Иртыша, Енисея, Лены, Амура. Особенно заметна их роль в обеспечении районов пионерного освоения, где практически отсутствуют сухопутные транспортные магистрали. В настоящее время в связи с экономическим кризисом идет сокращение объемов перевозок грузов и пассажиров речным транспортом, протяженности внутренних водных судоходных путей, уменьшается количество причалов.

Волжско-Камский бассейн, обслуживающий экономически наиболее развитые и плотно заселенные районы европейской части России, — главный. На него приходится свыше 1/2 грузооборота всего речного транспорта страны. Подавляющая часть перевозок в этом бассейне осуществляется по Волге, Каме и каналу имени Москвы. Наиболее крупными портами бассейна являются: три московских (Южный, Западный и Северный), Нижегородский, Казанский, Самарский, Волгоградский и Астраханский. [1.2]

На втором месте по объему выполняемой работы стоит Западно-Сибирский бассейн, включающий Обь с притоками. Здесь крупными портами выступают Новосибирск, Омск, Томск, Тобольск, Тюмень, Сургут, Уренгой, Лабытнанги.

Третьим по важности является воднотранспортный бассейн европейского Севера. Главной магистралью бассейна является Северная Двина с притоками Сухоной и Вычегдой. Ведущим портом бассейна выступает Архангельск.

Важное значение для снабжения Якутска и промышленных очагов Якутии играют Лена и расположенный на ее пересечении с БАМом порт Осетрово.

Стержнем воднотранспортной системы является Единая глубоководная система европейской части России общей протяженностью 6,3 тысячи километров. В нее входят глубоководные участки Волги (от Твери до Астрахани), Камы (от Соликамска до устья), Москвы-реки, Дона и межбассейновые глубоководные соединения — Московско-Волжское, Волго-Балтийское, Беломорско-Балтийское, Волго-Донское. Составляя лишь 6 % от общей протяженности внутренних водных путей, данная система выполняет свыше 2/3 всей перевозочной работы речного транспорта страны. На водных путях Единой глубоководной системы обеспечиваются гарантированные глубины до 4-4,5 м. [1.3]

Рассмотрим пути развития для речного транспорта в России. Основная доля зависит от:

* транспортной стратегии России на период до 2030 года;
* концепции развития внутреннего водного транспорта Российской Федерации на период до 2015 года, которая определяет основные направления развития внутреннего водного транспорта, а также цели, задачи и методы государственного регулирования в этой сфере в условиях развития рыночных отношений и структурного реформирования экономики.

Транспортная стратегия до 2030 года и концепции развития внутреннего водного транспорта до 2015 года показывают, что развитие речного транспорта в России напрямую зависит от стратегических документов, определяющих направления развития и методы государственного регулирования в этой сфере.

Важно учитывать современные вызовы, такие как рыночные отношения и изменения в экономике, чтобы адаптировать отрасль к новым условиям. Государственное регулирование должно способствовать стимулированию развития, поддержке инноваций, обеспечению безопасности и экологической устойчивости. [1.1]

Изучение методов кластеризации и решения поможет выявить факторы, влияющие на развитие речного транспорта, и определить конкретные стратегии для достижения поставленных целей. Например, кластеризация может позволить выделить группы судов по их техническим характеристикам, размерам, грузоподъемности и даже по экологической эффективности.

Для эффективного определения факторов, влияющих на развитие речного транспорта, и разработки конкретных стратегий необходимо провести анализ методов кластеризации и их решений. Кластеризация играет важную роль в анализе данных, однако, не все методы и алгоритмы могут эффективно обрабатывать большие массивы информации. Именно поэтому дальнейшие исследования направлены на преодоление этой проблемы.

1.2 Задача кластеризации и методы решения

Кластерный анализ является важным инструментом в области машинного обучения и статистики, позволяющим выделить скрытые закономерности и группировать данные для более глубокого понимания структуры набора данных. Он применяется в различных областях, включая маркетинговые исследования, медицинскую диагностику, обработку изображений и многие другие.

Критерии, на основе которых формируются кластеры, могут быть разнообразными: от числовых параметров, таких как размер или вес, до категориальных или даже текстовых характеристик объектов. Это позволяет проводить анализ данных различных типов и структур, позволяя делать более обоснованные выводы на основе группировки объектов.

Одним из ключевых преимуществ кластерного анализа является его способность обнаруживать неявные или сложные взаимосвязи между данными, выявляя группы, которые могут быть скрыты при первичном анализе. Это позволяет исследователям и аналитикам выделить подгруппы объектов, которые могут иметь схожие характеристики или поведенческие особенности, что дает возможность более точного понимания структуры данных.

Применение кластерного анализа в общем виде сводится к следующим этапам:

* отбор выборки объектов для кластеризации;
* определение множества переменных, по которым будут оцениваться объекты в выборке;
* вычисление значений меры сходства между объектами;
* применение метода кластерного анализа для создания групп сходных объектов (кластеров);
* представление результатов анализа.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата. [1.4]

Итак, как же определять «похожесть» объектов. Для начала нужно составить вектор характеристик для каждого объекта — как правило, это набор числовых значений, например, рост или вес человека. Однако существуют также алгоритмы, работающие с качественными характеристиками.

После того, как мы определили вектор характеристик, можно провести нормализацию, чтобы все компоненты давали одинаковый вклад при расчете «расстояния». В процессе нормализации все значения приводятся к некоторому диапазону.

Осуществив кластеризацию, возможно выявить отдельные данные, не относящиеся ни к одной группе. Чтобы определить наличие ошибки или интересного феномена, их следует обработать.

Иногда при сборе информации какие-то данные имеют больше признаков, а какие-то меньше. Изучаемый подход позволит сделать предположение об отсутствующих у иных элементов чертах группы. Приведем пример кластеризации.

Известно, что время нахождения участников в группе «мужчины» на сайте в среднем занимает 15 минут. Если в кластере появится новый человек с неизвестным временем пребывания на сайте, можно будет предположить, что для него оно тоже равно 15 минутам.

Различают много методов кластеризации, при использовании разных способов получаются свои результаты. Изучим самые распространенные методы для решения реальных задач. Первый — K-Means, мы рассмотрим его более подробно, так как он применяется в практической части. Чтобы использовать метод, надо придерживаться определенного алгоритма кластеризации.

Алгоритм K-Means включает следующие шаги:

1. Инициализация центроидов:

Начинаем с выбора k точек-центроидов из имеющихся данных.

1. Определение ближайших центроидов:

Для каждой точки данных вычисляем расстояние до всех центроидов, используя, например, евклидово расстояние:

где — значение признака *p* для точки данных — значение признака *p* для центроида .

Каждая точка данных относится к кластеру с ближайшим центроидом.

1. Пересчет центроидов:

Для каждого кластера находим новый центр, который является средним значением координат всех точек в кластере:

где — набор всех точек данных, отнесенных к определенному кластеру *j*, а — количество точек в кластере.

Эти шаги повторяются до тех пор, пока центроиды кластеров не стабилизируются или не будет достигнуто максимальное количество итераций.

Следующий метод, который мы обсудим, — это DBSCAN. В отличие от предыдущих методов, количество кластеров определяется алгоритмом автоматически. Однако для успешной реализации DBSCAN крайне важно определить два параметра: диапазон поиска и минимальное количество точек, которое необходимо: [1.7]

* осуществим выбор объекта и найдем точки вокруг него в выбранном нами диапазоне;
* если их не набралось даже минимального количества, назовем их выбросами и не будем относить ни к одной группе;
* если удалось найти нужное их количество, для каждой также ищем новые точки в выбранном нами диапазоне. Так, все из них, которые находятся друг от друга на расстоянии меньшем или равном заданному диапазону, будут формировать один кластер.

Однако выбор подхода к использованию метода не всегда очевиден. В зависимости от задачи и особенностей данных, различные способы кластеризации более или менее эффективны. Метод DBSCAN особенно полезен, когда данные содержат объекты, не принадлежащие ни к одной группе, и кластеры имеют сложную форму.

Определение количества классов K-Means является одним из самых популярных методов кластеризации и обычно используется, когда заранее известно их количество, которое нужно обнаружить в данных. Он хорошо работает с линейно разделимыми данными, когда можно провести линии, разделяющие кластеры.

Если задача требует определения точного количества групп, K-Means может быть более подходящим. Но если данные не могут быть разделены линейными линиями, то метод DBSCAN имеет преимущества. Изображение, представленное ниже, свидетельствует о том, что DBSCAN точнее определяет кластеры тогда, когда данные неразделимы таким образом, чем K-Means.

Оценка результатов кластеризации является сложной задачей, так же как и сам процесс группировки. Наиболее распространенные методы включают «внутреннюю» и «внешнюю» оценку.

В первом случае система сводится к одному качественному показателю, а во втором кластеризацию сравнивают с уже существующей классификацией или «основополагающей правдой». Дополнительно эксперт-человек может провести ручную оценку и определить полезность использования метода в предполагаемом приложении.

Кластеризация — очень полезный инструмент, особенно в области анализа данных для рекламы. Когда необходимо эффективно распределить PR-бюджет, привлекая максимальное количество клиентов за минимальные затраты, метод поможет определить наиболее подходящий подход.

Исследование методов кластеризации для анализа развития речного транспорта подчеркивает их значимость в обработке данных для выявления факторов и стратегий развития. Оценка результатов кластеризации представляет собой сложную задачу, требующую как объективной, так и субъективной оценки. Различные подходы к оценке, как внутренние, так и внешние, позволяют оценивать качество формирования кластеров.

Для практической части исследования следует осуществить выбор метода кластеризации, который наилучшим образом подходит для анализа данных в области речного транспорта. Затем необходимо применить этот метод к имеющимся данным, анализируя и интерпретируя полученные кластеры с целью выявления факторов, влияющих на развитие речного транспорта.

Дополнительно, для оценки результатов, можно провести сравнение полученных кластеров с уже существующими классификациями или экспертными оценками.

При выборе метода кластеризации важно учитывать специфику данных речного транспорта, такие как объем перевозок, частота рейсов, маршруты и типы судов. Применение методов, таких как иерархическая кластеризация, метод K-средних или DBSCAN, может дать различные перспективы на структуру данных и выявить скрытые паттерны.

2 ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1 Загрузка данных и определение количества кластеров

В данной главе представлен анализ кластеризации данных о лодках, включая характеристики, такие как год постройки, длина, ширина и количество просмотров за последние 7 дней. Целью данного исследования является оценка эффективности метода кластеризации для определения оптимального количества кластеров и анализа характеристик этих кластеров.

Для достижения этой цели использовались различные методы, включая загрузку и предобработку данных, определение оптимального количества кластеров с помощью метода "Elbow", кластеризацию методом k-средних, а также визуализацию и оценку качества кластеров.

Данные были загружены из файла "boat\_data.csv" и предварительно обработаны: переменные были приведены к корректным именам, а нулевые значения в столбце "Year.Built" заменены на NA, после чего данные с пропущенными значениями были удалены. Далее, данные были стандартизированы для обеспечения корректного применения методов кластеризации.

Для определения оптимального количества кластеров был использован метод "Elbow". В рамках этого метода рассчитывалась сумма квадратов внутри групп (wss) для различного количества кластеров (от 1 до 15). Затем был построен график зависимости суммы квадратов внутри групп от количества кластеров, представленный на Рисунке 2.1. На основе графика был сделан вывод, что оптимальное количество кластеров составляет 5. Это значение было выбрано, так как на графике наблюдается явный "локоть" в этой точке, что указывает на значительное снижение суммы квадратов внутри групп при увеличении количества кластеров до 5.

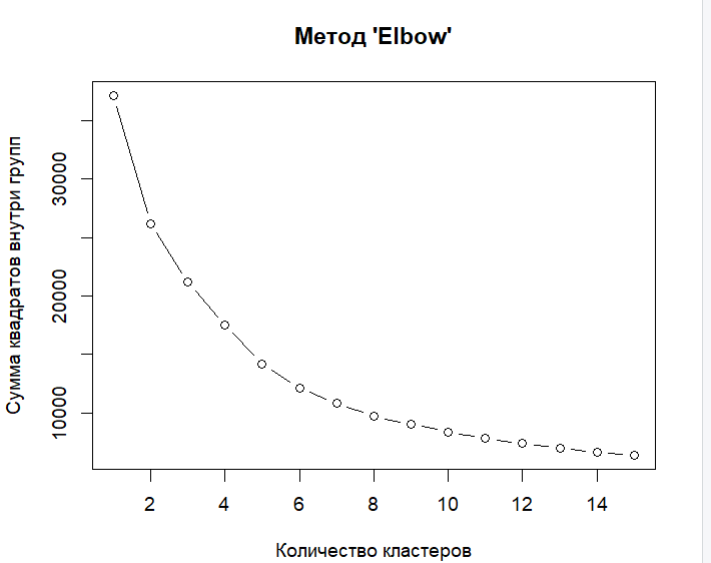


Рисунок 2.1 — График Elbow

На данном графике видно, что при увеличении количества кластеров от 1 до 5 наблюдается резкое снижение суммы квадратов внутри групп, после чего снижение становится менее заметным. Таким образом, оптимальным количеством кластеров является пять, что и было использовано в дальнейшем анализе.

2.2 Кластеризация и анализ результатов

Во второй части работы проводилась кластеризация методом k-средних, визуализировались результаты кластеризации и оценивалось качество кластеров с использованием различных метрик. Также проводился анализ характеристик кластеров с помощью радар-чартов.

Для проведения кластеризации методом k-средних было выбрано 5 кластеров. Этот метод был реализован с использованием функции kmeans(), при этом параметр nstart был установлен на 25 для повышения стабильности результатов. Результаты кластеризации представлены на графике на Рисунке 2.2.

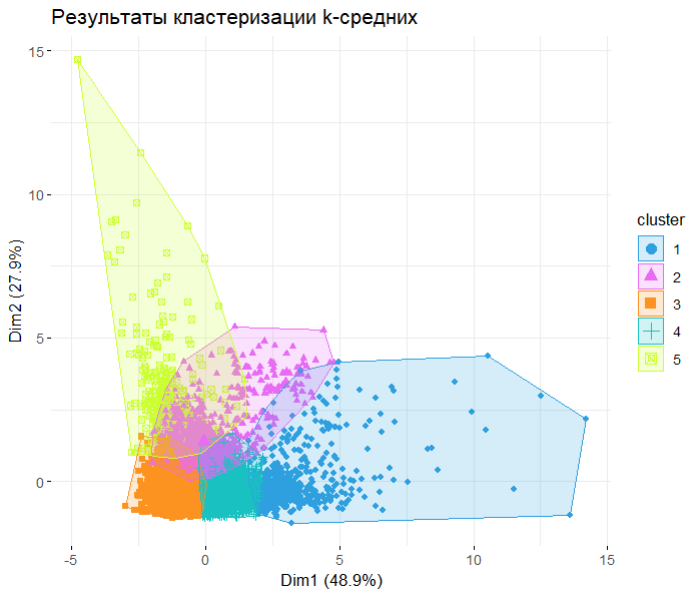


Рисунок 2.2 — Визуализация результатов кластеризации

График показывает результаты кластеризации методом k-means, где объекты распределены по двум измерениям (Dim1 и Dim2). Каждый кластер обозначен различным цветом и символом. Оси X и Y представляют собой первые два главных компонента (Dim1 и Dim2), которые объясняют 48.9 % и 27.9 % вариации данных соответственно.

Описание кластеров:

1. Кластер 1 (Голубой, ромбы):

Значительная часть данных попала в этот кластер, который имеет широкую распределенность по оси X (Dim1). Объекты этого кластера более плотно сгруппированы, чем в других кластерах, что указывает на хорошую однородность.

1. Кластер 2 (Розовый, треугольники):

Этот кластер имеет более компактное распределение по обеим осям, сосредоточенное в центре графика. Взаимное перекрытие с кластером 5 указывает на возможность некорректного разделения или наличие пограничных объектов.

1. Кластер 3 (Оранжевый, квадраты):

Этот кластер представлен меньшим числом объектов и расположен в нижнем левом углу графика. Наблюдается плотная концентрация объектов, что может говорить о хорошей кластеризации, однако малое количество объектов может быть недостатком.

1. Кластер 4 (Бирюзовый, кресты):

Расположен ниже и ближе к центру графика. Несмотря на компактное распределение, есть некоторое пересечение с кластером 1, что может указывать на пограничные объекты.

1. Кластер 5 (Желтый, квадраты):

Имеет широкое распределение по оси Y (Dim2) и занимает значительную часть графика. Имеет пересечение с кластерами 2 и 4, что также указывает на возможные проблемы с границами кластеров.

После проведения кластеризации мы проанализируем график силуэта кластеров, представленный на Рисунке 2.3, для оценки качества полученных кластеров. График силуэта кластеров предоставляет информацию о том, насколько объекты внутри кластеров похожи друг на друга и насколько они отличаются от объектов в других кластерах.

На графике силуэта кластеров ось Y представляет ширину силуэта (Silhouette width Si), которая измеряет, насколько объект хорошо согласуется со своим собственным кластером по сравнению с другими кластерами. Значения силуэта варьируются от -1 до 1: значения ближе к 1 указывают на хорошо классифицированные объекты, значения около 0 говорят о неопределенности (пограничных объектах), а отрицательные значения означают, что объекты, возможно, были неправильно классифицированы.

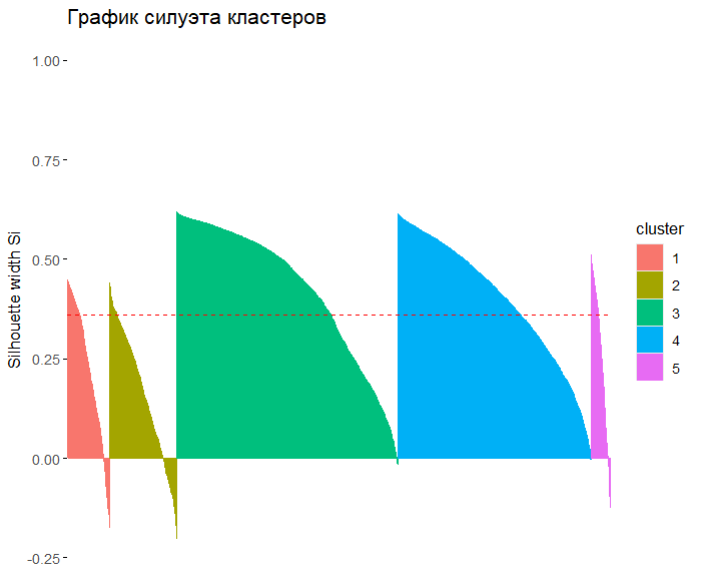


Рисунок 2.3 — График силуэта кластеров

Ось X представляет собой кластеры, на которые разбиты данные. Каждый кластер обозначен на графике, и высота столбца силуэта указывает на среднюю ширину силуэта для объектов в соответствующем кластере.

Далее мы проанализировали значения силуэта для каждого кластера и сделали следующие выводы:

1. Кластер 1 (Красный):

Средняя ширина силуэта для объектов этого кластера находится в диапазоне около 0.15, с некоторыми значениями, опускающимися до -0.25. Наличие отрицательных значений указывает на возможные ошибки классификации для некоторых объектов внутри этого кластера. В целом, этот кластер имеет относительно низкие значения силуэта, что говорит о плохом качестве кластеризации.

1. Кластер 2 (Оливковый):

Объекты в этом кластере имеют ширину силуэта, в основном, около 0.05. Очень низкие значения ширины силуэта указывают на неопределенность и возможные проблемы с правильностью кластеризации.

1. Кластер 3 (Зеленый):

Этот кластер демонстрирует высокие значения ширины силуэта, в основном в диапазоне 0.3-0.55. Высокие значения силуэта указывают на хорошую кластеризацию объектов внутри этого кластера.

1. Кластер 4 (Голубой):

Объекты этого кластера имеют высокие значения ширины силуэта, варьирующиеся от 0.35 до 0.5. Как и в кластере 3, высокие значения силуэта указывают на хорошее качество кластеризации.

1. Кластер 5 (Розовый):

Значения ширины силуэта для этого кластера колеблются около 0, с несколькими значениями около -0.25. Это указывает на то, что некоторые объекты могут быть неправильно классифицированы, и общий уровень неопределенности в этом кластере выше.

Исходя из анализа графика силуэта кластеров, мы можем сделать следующие выводы:

1. Кластеры с высокими значениями ширины силуэта (кластеры 3 и 4) демонстрируют хорошую однородность объектов внутри себя и высокое качество кластеризации. Это свидетельствует о том, что объекты в этих кластерах сильно схожи друг с другом и отличаются от объектов в других кластерах.
2. Кластеры с низкими или отрицательными значениями ширины силуэта (кластеры 1, 2 и 5) указывают на неопределенность и возможные ошибки в кластеризации. Это может быть связано с наличием пограничных объектов, которые могут быть неправильно классифицированы, или с недостаточной однородностью внутри кластеров.

Помимо анализа графика силуэта кластеров, была проведена радарная диаграмма для представления характеристик различных кластеров. Каждый кластер обозначен уникальной цветовой линией и стилем, а параметры представлены в нормированном виде для сравнения значений в диапазоне от 0 до 100. График радарной диаграммы представлен на Рисунке 2.4.

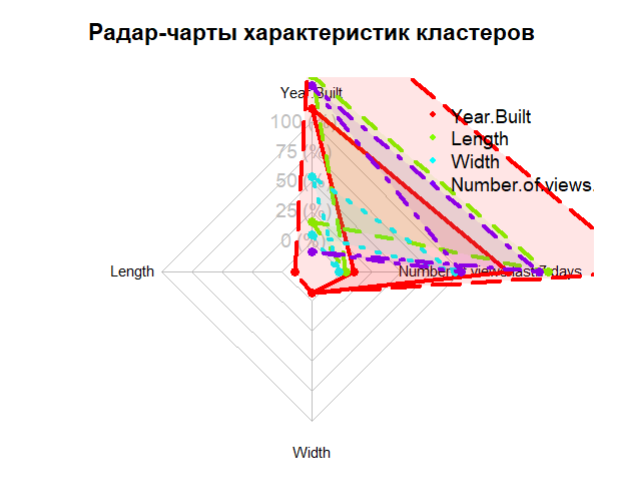


Рисунок 2.4 — Радарная диаграмма

Давайте подробнее рассмотрим, что изображено на данной диаграмме.

Ось "Year Built" (Год постройки): Эта ось отражает год, в котором была построена лодка. Чем выше значение на этой оси, тем более новая лодка.

Ось "Length" (Длина): Показывает длину лодки в метрах. Большие значения на этой оси указывают на более длинные лодки.

Ось "Width" (Ширина): Отображает ширину лодки в метрах. Широкие лодки имеют большие значения на этой оси, что может свидетельствовать о большей устойчивости и вместительности.

Ось "Number of Views" (Количество просмотров): Показывает, сколько раз лодку просмотрели. Высокие значения могут указывать на популярность лодки или активность её просмотров.

Каждый кластер на диаграмме представляет собой комбинацию этих характеристик для разных лодок. Например, если кластер высоко оценен по осям "Length" и "Width", это может указывать на большую и просторную лодку. Если кластер имеет высокие значения по оси "Year Built", это говорит о том, что лодки в этом кластере относительно новые. А высокие значения на оси "Number of Views" могут свидетельствовать о высоком интересе к лодкам этого кластера.

Далее следует перейти к рассмотрению статистики, а именно различных численных коэффициентов. После проведения кластеризации лодок и оценки качества разделения на кластеры получены следующие метрики:

1. Количество кластеров: 5
2. Средняя ширина силуэта по кластерам: 0.361
3. Индекс Данна: 0.001
4. Индекс Данна 2: 0.716
5. Критерий Калински-Харабасза: 3736.773
6. Индекс Сокала-Снида: 0.172

На основе проведённого анализа кластеризации лодок, можно сделать следующие выводы. Высокие значения средней ширины силуэта (0.361) и критерия Калински-Харабасза (3736.773), а также индекса Данна 2 (0.716) указывают на хорошее качество кластеризации с чётким разделением и высокой схожестью объектов внутри кластеров. Тем не менее, низкие значения индекса Данна (0.001) и индекса Сокала-Снида (0.172) свидетельствуют о возможных проблемах с разделением на кластеры и наличии пограничных объектов. Это может потребовать дальнейшей настройки модели кластеризации или дополнительного анализа данных.

ЗАКЛЮЧЕНИЕ

Данное исследование направлено на анализ динамичной и конкурентной отрасли рынка лодок и катеров. Цель данного исследования — применение методов анализа и прогнозирования для создания модели, способной разделить лодки на кластеры на основе их характеристик и размеров, для выявления групп схожих судов — достигнута. Для достижения этой цели мы изучили научную и методическую литературу, связанную с анализом рынка водных средств транспорта, а также методами кластерного анализа.

Мы собрали необходимые данные о лодках и катерах, включая их размеры, характеристики и данные о просмотрах объявлений. С использованием языка программирования R была проведена статистическая обработка данных и работа с графикой для кластеризации лодок на основе их размеров и других характеристик.

В результате были созданы модели прогнозирования разделения лодок на кластеры, а их результаты подверглись качественному анализу. В ходе работы были выполнены следующие задачи:

* изучена научная и методическая литературу, связанную с анализом рынка водных средств транспорта, а также методами кластерного анализа;
* собраны необходимые данные о лодках и катерах, включая их размеры, характеристики и данные о просмотрах объявлений;
* применен язык программирования R для статистической обработки данных и работы с графикой для проведения кластеризации лодок на основе их размеров и других характеристик;
* созданы модели прогнозирования разделения лодок на кластеры и качественный анализ результатов этой модели;

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

* 1. Давыдов, В. В. Технические вычисления в кораблестроении / В.В. Давыдов. — М.: Речной транспорт, 2000. - 248 c. (Дата обращения: 07.05.2024).
  2. Дмитриев, В. И. Пути повышения безопасности судоходства. Учебное пособие: моногр. / В.И. Дмитриев. — М.: Моркнига, 2015. - 224 c. (Дата обращения: 01.05.2024).
  3. Водный транспорт [Электронный ресурс]. — URL: <https://studopedia.ru/19_97111_vodopriemniy-kovsh-neobhodimo-> (Дата обращения: 15.05.2024).
  4. EМ — масштабируемый алгоритм кластеризации [Электронный ресурс]. - URL: <https://loginom.ru/blog/em> (Дата обращения: 02.05.2024).
  5. Слуцкин, Л. Н. Анализ стабильности модели линейной регрессии во времени / Л.Н. Слуцкин. — М.: Синергия, 2007. - 283 c. (Дата обращения: 07.05.2024).
  6. Выгодчикова, И. Ю. Алгоритм оценки параметров кластерной модели по минимаксному критерию / И.Ю. Выгодчикова. — М.: Синергия, 2015. - 125 c. (Дата обращения: 07.05.2024).
  7. Метод кластеризации / Студенческие реферативные статьи и материалы "Studref" Режим доступа: [https://studref.com](https://studref.com/591347/ekonomika/metod). (Дата обращения: 07.05.2024).
  8. Герасимова, Е. А. Эконометрика: регрессионный анализ : учебно-методическое пособие / Е. А. Герасимова, М. Ю. Карышев. — Самара: СамГУПС, 2016. — 98 с. (Дата обращения: 08.05.2024).
  9. Бергер Е.Г. Нормоконтроль документации [Электронный ресурс]: Методические рекомендации / Бергер Е.Г., Зуев А.С. — М.: МИРЭА — Российский технологический университет, 2020. — 1 электрон. опт. диск (CD-ROM). Дата обращения (07.05.2024).

ПРАКТИЧЕСКАЯ ЧАСТЬ

1. Исходные данные [Электронный ресурс] — URL: <https://www.kaggle.com/datasets/karthikbhandary2/boat-sales> (Дата обращения: 01.05.2024).
2. Алгоритм кластеризации [Электронный ресурс]. — URL: <https://loginom.ru/blog/em> (Дата обращения: 11.05.2024).
3. Оценка параметров кластеризации [Электронный ресурс]. — URL: <https://studfile.net/preview/2014530/page:3/> (Дата обращения: 07.05.2024).
4. Метрики качества кластерных моделей [Электронный ресурс]. — URL: <https://loginom.ru/blog/quality-metrics> (Дата обращения: 07.05.2024).

ПРИЛОЖЕНИЕ

Приложение А —Программный код.

Приложение А

В Листинге А.1 представлен итоговый программный код проекта на языке R.

Листинг А.1 — Программный код

# Загрузка библиотек

library(cluster)

library(factoextra)

library(vegan)

library(ggplot2)

# Загрузка и предобработка данных

data <- read.csv("boat\_data.csv", stringsAsFactors = FALSE)

names(data) <- make.names(names(data), unique = TRUE)

print(names(data))

data$Year.Built <- ifelse(data$Year.Built == 0, NA, data$Year.Built)

data <- na.omit(data)

data\_scaled <- scale(data[, c("Year.Built", "Length", "Width", "Number.of.views.last.7.days")])

head(data\_scaled)

# Определение количества кластеров методом "Elbow"

wss <- (nrow(data\_scaled) - 1) \* sum(apply(data\_scaled, 2, var)) # Вычисление общей суммы квадратов внутри групп

for (i in 2:15) wss[i] <- sum(kmeans(data\_scaled, centers = i, nstart = 25)$withinss) # Цикл для вычисления суммы квадратов внутри групп для разного количества кластеров

plot(1:15, wss, type = "b", xlab = "Количество кластеров", ylab = "Сумма квадратов внутри групп", main = "Метод 'Elbow'") # Построение графика "Elbow method"

# Кластеризация k-средних

set.seed(123)

kmeans\_result <- kmeans(data\_scaled, centers = 5, nstart = 25)

print(kmeans\_result)

# Визуализация результатов кластеризации

fviz\_cluster(kmeans\_result, data = data\_scaled, geom = "point", stand = FALSE,

ellipse = TRUE, show.clust.cent = TRUE,

palette = c("#2E9FDF", "#E76BF3", "#FC921F", "#19C1C1", "#CCFF33"),

ggtheme = theme\_minimal(),

main = "Результаты кластеризации k-средних")

#Это уже для курсовой

install.packages("fpc")

library(fpc)

Продолжение листинга А.1

# Calculate Davies-Bouldin Index

cluster\_stats <- cluster.stats(dist(data\_scaled), kmeans\_result$cluster)

# Вывод интересных статистик кластеров

print("Cluster Quality Metrics:")

print(paste("Number of clusters:", cluster\_stats$cluster.number))

print(paste("Average silhouette width across clusters:", cluster\_stats$avg.silwidth))

print(paste("Dunn index:", cluster\_stats$dunn))

print(paste("Dunn2 index:", cluster\_stats$dunn2))

print(paste("Calinski-Harabasz criterion value:", cluster\_stats$ch))

print(paste("Sokal-Sneath index:", cluster\_stats$sindex))

# Print cluster statistics

print(cluster\_stats)

# Визуализация силуэта

silhouette\_scores <- silhouette(kmeans\_result$cluster, dist(data\_scaled))

fviz\_silhouette(silhouette\_scores, main = "График силуэта кластеров")

# Радар-чарты для анализа характеристик кластеров

# Загрузка библиотеки fmsb

library(fmsb)

# Радар-чарты для анализа характеристик кластеров

data\_mean <- aggregate(data\_scaled, by=list(cluster=kmeans\_result$cluster), FUN=mean) # Вычисление средних значений переменных по кластерам

data\_mean <- data\_mean[, -1] # Удалить колонку кластера

radar\_data <- rbind(rep(max(data\_mean), ncol(data\_mean)), data\_mean, rep(min(data\_mean), ncol(data\_mean)))

colnames(radar\_data) <- colnames(data\_mean)

radarchart(radar\_data, axistype=1,

pcol=rainbow(ncol(data\_mean)), pfcol=rainbow(ncol(data\_mean), alpha=0.1),

plwd=4, cglcol="grey", cglty=1, axislabcol="grey",

vlcex=0.8, title = "Радар-чарты характеристик кластеров")

# Добавление легенды

legend(x=0.7, y=1.2, legend=colnames(data\_mean), bty="n", pch=20, col=rainbow(ncol(data\_mean)))