

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ  
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ  
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ  
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ  
УНИВЕРСИТЕТ»  
(НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Факультет экономический

Кафедра применения математических методов в экономике и планировании

Направление подготовки 38.03.01 Экономика

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

Балабаев Владислав Дмитриевич

(Фамилия, Имя, Отчество автора)

Тема работы Эмпирическая оценка эффекта гарантий на рынке легковых автомобилей России

**«К защите допущен»**

Заведующий кафедрой,  
д. э. н., профессор,

**Научный руководитель**

к.э.н., доцент,  
ЭФ НГУ

Мкртчян Г. М./\_\_\_\_\_

«\_\_\_\_ » \_\_\_\_\_ 2023г.

Шильцин Е. А./\_\_\_\_\_

«\_\_\_\_ » \_\_\_\_\_ 2023г.

Дата защиты «\_\_\_\_ » \_\_\_\_\_ 2023г.

Новосибирск  
2023

## ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ.....	3
ГЛАВА 1. ХАРАКТЕРИСТИКА АВТОМОБИЛЬНОГО РЫНКА РОССИИ И ОПИСАТЕЛЬНЫЙ АНАЛИЗ ИМЕЮЩИХСЯ ДАННЫХ.....	5
1.1. Краткое описание рынка автомобилей РФ.....	5
1.2. Исследование собранных данных объявлений с площадки Drom.ru и их описательный анализ.....	7
ГЛАВА 2. ВЫБОР ЭКОНОМЕТРИЧЕСКИХ ПОДХОДОВ И МЕТОДОВ И ИХ ТЕОРЕТИЧЕСКАЯ ПОСТАНОВКА.....	22
2.1. Модели и методы машинного обучения, применяемые для кластеризации.....	22
2.2. Эконометрические методы пространственного анализа.....	28
2.3. Алгоритм определения эффекта гарантий.....	32
ГЛАВА 3. ОПИСАНИЕ И ОЦЕНКА ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ.....	37
3.1. Краткая характеристика кластерного анализа.....	37
3.2. Оценка результатов пространственного анализа.....	40
3.3. Эмпирическое оценивание эффекта гарантий, интерпретация.....	41
ЗАКЛЮЧЕНИЕ.....	45
СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ.....	47
ПРИЛОЖЕНИЕ А. Оценки модели SLM с фиксированными эффектами для различных ценовых категорий.....	51

## **ВВЕДЕНИЕ**

Актуальность данной работы во многом обуславливается динамичностью развития автомобильного рынка Российской Федерации с 2021 года по настоящее время.

После ухода большинства игроков автомобильной отрасли из России [BMW Halts Production..., 2022; Volkswagen в РФ..., 2022], приостановления экспортных поставок [Компания Rolls-Royce..., 2022] и санкционного запрета на ввоз автомобилей [Imposition of Sanctions..., 2022; Ukraine: EU..., 2022], автомобильный рынок России изменяется не только количественно, но и структурно.

Данные изменения, помимо прочего, приводят к увеличению доли машин с пробегом, возрастают количество официально работающих компаний, занимающихся поставками и торговлей вторичных автомобилей, и их обороты, а также, увеличивается количество частных лиц и организаций, занимающихся данной деятельностью в теневом секторе.

Присутствие теневого сегмента на автомобильном рынке России в первую очередь обуславливается налогообложением, вместе с этим, именно благодаря уклонению от налогообложения неофициально работающие физические лица и компании способны конкурировать с легальными коммерческими организациями, неся меньшие издержки и предоставляя более низкие цены.

Однако, вопреки более высоким ценам предложения, наличие юридических предприятий на вторичном сегменте данного рынка, помимо очевидной законности ведения бизнеса, во многом сводиться к гарантиям, которые они способны предоставить, что включает ответственность, взятую на себя.

Эффективная количественная оценка эффекта гарантит может служить качественным показателем для построения оптимальной налоговой политики, которая будет способствовать тому, что ведение теневой деятельности на вторичном автомобильном рынке будет экономически нецелесообразным, при этом не приводя сильному уменьшению налоговых поступлений, в результате уменьшения налоговых ставок.

Целью работы является анализ автомобильного рынка Российской Федерации с осуществлением эмпирической оценки ценообразующих факторов и эффекта гарантит при формировании цены на новые и вторичные автомобили.

Определение поставленной цели позволяет выделить определенный перечень задач:

1. Охарактеризовать текущее состояние автомобильного рынка РФ;
2. Провести первичный анализ макроэкономических данных автомобильного рынка России;

3. Сформировать перечень данных, необходимых для исполнения анализа объекта исследования;
4. Провести описательный анализ имеющихся данных, сформировать гипотезы и выводы;
5. Рассмотреть подходы, которые позволяют провести эффективную оценку имеющихся данных;
6. Осуществить эмпирический анализ имеющихся данных на основе предложенных моделей;
7. Провести оценку полученных результатов.

Объект исследования: Автомобильный рынок РФ, характеристики предложения.

Предмет исследования: Эмпирическая оценка ценообразующих факторов на автомобильном рынке РФ.

# **ГЛАВА 1. ХАРАКТЕРИСТИКА АВТОМОБИЛЬНОГО РЫНКА РОССИИ И ОПИСАТЕЛЬНЫЙ АНАЛИЗ ИМЕЮЩИХСЯ ДАННЫХ**

## **1.1. Краткое описание рынка автомобилей РФ**

За 2022 год продажи новых автомобилей на территории Российской Федерации снизились более, чем на 59%, составив 687 тыс. автомобилей [Статистика продаж..., 2023], при этом рост цен на новые легковые автомобили составил 24%, где сегмент иномарок подорожал более чем на 35% [Индексы потребительских..., 2023]. Данные показатели являются наихудшими за последние 20 лет. Самое сильное повышение цен произошло в марте — в среднем на 23%, однако, на отдельные модели цены увеличивались до 60%.

Относительно начала 2022 года, число объявлений на сайте-агрегаторе автомобильных объявлений “Авто.ру” упало вдвое к концу данного года [Как автобизнес..., 2022]. В результате изменения цен, большим изменениям подверглась и сама структура рынка автомобилей. Доля новых автомобилей среднего ценового сегмента уменьшилась на 50%, тогда как доля автомобилей с ценой выше 3 млн. руб. увеличилась вдвое. Основную долю рынка на конец 2022 года занимают машины в ценовом сегменте от 3 до 5 млн. руб. — 35%.

В результате, рынок новых автомобилей испытывает дефицит в категории новых автомобилей стоимостью от 800 тыс. до 1,5 млн. руб., что создает повышенный спрос на вторичном рынке.

Подсчитать долю нелегально работающих профессиональных частных продавцов подержанных автомобилей представляется сложной задачей, как и любая оценка теневого сегмента любого рынка, поэтому, общедоступной официальной информации по данному показателю от какого-либо статистического агрегатора нет. Вместе с этим, разные не статистические информационные источники сообщают о разных интервалах доли данного сегмента продавцов, поэтому, составление консенсусного мнения о приблизительном значении данного показателя среди них тоже представляется невозможным.

Однако, нельзя отрицать факт наличия таких участников на данном рынке, роль которых в 2022 году существенно возросла [Рынок автомобилей..., 2022] по причине разрыва сложившейся за многие годы логистики импорта новых и подержанных автомобилей и прямых запретов на ввоз иномарок.

Наличие физических лиц, которые нелегально занимаются профессиональной перепродажей автомобилей и работающих в теневом сегменте, а также неофициально работающих салонов, во многом обуславливается неспособностью официально работающих

компаний конкурировать с ними, особенно, в низко бюджетном ценовом сегменте, по причине отсутствия достаточной маржинальности, которая следует из того, что они должны предоставлять сопоставимые с ФЛ цены при этом являясь налогоплательщиками.

Для того, чтобы частное лицо могло продать автомобиль салону или другому частному лицу, необходимо совершить ряд действий [Постановление Правительства..., 2019], а именно, заключить договор купли-продажи, ДКП, расписаться в ПТС, с 1 ноября 2020 года на все новые автомобили в салонах оформляют только электронный ПТС [Решение Коллегии..., 2020], ЕПТС, передать свидетельство о регистрации, а также подать декларацию о доходах, налог с которых до необходимо оплатить до 15 июля следующего года с момента продажи. При этом, новый собственник автомобиля обязан поставить его на учет в течение 10 суток после приобретения.

Доход при продаже автомобиля частным лицом, являющимся резидентом, облагается НДФЛ в размере 13%. Однако, участники теневого рынка в подавляющем большинстве случаев обходят выплату НДФЛ. При продаже ФЛ имущества учитывается так называемый минимальный срок владения – период, по истечении которого собственник не обязан подавать декларацию и платить НДФЛ, который для автомобилей составляет 3 года [Доходы, не подлежащие..., 2023].

Поэтому, во избежание налогообложения, большинство сделок по неподтвержденной неофициальной информации совершается теневыми участниками рынка следующим образом. При осуществлении сделки с собственником автомобиля полностью заполняется ДКП, кроме лица, которому был продан автомобиль, в дальнейшем, при осуществлении продажи автомобиля найденный покупатель вписывается в ДКП. Данная операция должна быть совершена в течение 10 суток, для того, чтобы успеть поставить автомобиль на учет. В результате, теневой участник рынка юридически нигде не значится в обеих сделках и его доход не налогооблагается.

Помимо минимального срока владения существует также максимальный налоговый вычет, который в России для имущества, кроме недвижимого, равняется 250 тыс. руб. [Имущественные налоговые..., 2023]. Это значит, что если автомобиль был продан не более чем за 250 тыс. руб., платить НДФЛ с него не нужно.

Если автомобиль продается официально работающей компанией или индивидуальным предпринимателем, то уплата налогов при продаже автомобиля сводиться к НДС и налогу на прибыль для юридических лиц, НДФЛ в размере 13% для ИП на ОСН или НДП, и 15% налог для ИП на УСН «Доходы минус расходы». Вместе с этим, стоит учитывать также НДФЛ и взносы, которые платятся юридическим лицом или ИП, имеющим работников по трудовому

договору, что тоже увеличивает налоговую нагрузку по сравнению с компанией, работающей в теневом секторе.

В результате, неофициально работающие участники данного рынка действительно имеют некоторые выгоды, однако, основной фактор, помимо масштабируемости, связанной с легальностью бизнеса, который позволяет официальным автосалонам конкурировать с теневым сегментом – гарантии, которые они предоставляют покупателям и продавцам автомобилей, и которые закладываются в цену автомобиля.

Однако, эффект гарантий не является настолько значительным, чтобы вывести теневых игроков из данного рыночного сегмента. Именно поэтому основной целью настоящей квалификационной работы является оценка данного эффекта гарантий на автомобильном рынке России, с помощью которой можно выстраивать налоговую политику, которая бы способствовала эффективному противоборству с теневым сегментом данного рынка.

Ожидается, что величина эффекта гарантий разнится для автомобилей, находящихся в различных ценовых сегментах, более того, эмпирические представления о потенциальном значении эффекта гарантий говорят о том, что эффект гарантий может различаться как и внутри ценового сегмента между определенными моделями. Однако, для того, чтобы получить более асимптотически близкие оценки, необходимо иметь достаточное количество данных, которые в дальнейшем будут соответствовать только для соответствующей состоятельной оценки величины эффекта гарантий для различных ценовых сегментов в целом.

## **1.2. Исследование собранных данных объявлений с площадки Drom.ru и их описательный анализ**

Для оценки эффекта гарантий необходимо иметь данные о конкретных автомобильных единицах, которые имеют определенные характеристики, а главное, цену. Общедоступная база микроданных по России, которые бы содержали наблюдения на уровне единичных автомобилей, является достаточно скучной и не имеет актуальной информации.

Самым крупным статистическим агентством России в автомобильной отрасли является аналитическое агентство, работающее с 2005 года, “АВТОСТАТ” [Официальный сайт... 2023], имеющее платную аналитическую систему Russian Automotive Data Analytics & Research, которая содержит полное историческое и пространственное представление об автомобилях, которые учитываются ГИБДД с помощью ЕПТС. Данная система была проанализирована как потенциальный источник для настоящей квалификационной работы, однако, основным минусом

данной базы данных является отсутствие ценовой привязки между наблюдениями, что не позволяет использовать приведенные данные.

Согласно статистике аналитического агентства “АВТОСТАТ”, наибольшую долю рынка платформ для покупки-продажи автомобилей на середину 2022 года [Авито Авто..., 2022] имеет “Авито Авто” – 56,6%, далее, второе и третье места делят “Auto.ru” и “Drom.ru”, имея по 21,2% и 20,1% соответственно. Данные площадки не предоставляют и не продают исторические данные выставленных объявлений коммерческим и некоммерческим организациям, обосновывая это внутренней политикой, а также соблюдением законодательства РФ о защите пользователей, предоставляя доступ только правоохранительным органам. Однако, имеется возможность самостоятельного сбора информации о текущих объявлениях с данных сайтов-агрегаторов.

Так как данные площадки не являются смещенными по каким либо категориям автомобилей, общность объявлений которых является достаточно близкой для того, чтобы сделать основательные выводы о генеральной выборке, при этом, объявления о продаже автомобилей достаточно часто дублируются на данных площадках, целесообразно сфокусироваться на исследовании объявлений одного из данных веб-агрегаторов. В силу сравнительного удобства использования данных был выбрана площадка Drom.ru.

Сайт Drom.ru ежедневно содержит более полумиллиона активных объявлений о продаже новых автомобилей, в том числе объявлений от дилеров, и автомобилей с пробегом. В каждом объявлении отображается цена автомобиля, его марка и модель, локация, в которой продается автомобиль, например, город, поселок или рынок, а также дата самого объявления и количество его просмотров пользователями. Помимо этих данных, все остальные данные, которые может содержать объявление, являются опциональными, в том числе и фотографии самого автомобиля.

У каждого объявления присутствуют отдельно отображаемые характеристики автомобиля, список которых, за исключением обязательных, выбирается самим создателем объявления и которые заполняются им же. Среди таких характеристик самыми основными являются следующие: тип кузова, тип двигателя и его объем, мощность, коробка передач, привод, цвет, руль – правый или левый, а также пробег автомобиля в километрах, с учетом информации о том, имеется ли пробег у автомобиля вообще или пробег в РФ в частности, также учитывается поколение автомобиля и его комплектация. Так как содержимое объявления вариативно, необходимо учитывать возможность того, что в очередном объявлении не будет какого-либо заполненного параметра автомобиля, а в каком-то будет.

Вместе с этим, некоторые объявления могут содержать информацию по VIN-коду, которая, например, говорит о том, совпадают ли указанные характеристики автомобиля с ПТС, сколько имеется записей о регистрации, числится ли автомобиль в розыске.

Также у некоторых объявлений присутствует оценка цены автомобиля алгоритмами самого сайта Drom.ru, которая может сообщить о том, является ли цена высокой, нормальной или низкой относительно тех характеристик автомобиля, которые были указаны. Практически каждое объявление содержит информацию о годе производства конкретного автомобиля.

Большое количество объявлений о продаже автомобилей на сайте не позволяет собрать данную информацию вручную, а сбор лишь выборочной части может оказаться в результате моделирования дезориентирующим и смешенным в категорию каких-либо конкретных марок и моделей автомобилей. Данный сайт-агрегатор объявлений о продаже автомобилей, а также сайты Avito.ru, Auto.ru, не позволяют выгружать данные непосредственно. Поэтому, для того, чтобы собрать описанные выше микро-данные необходимо было построить процесс автоматического сбора информации, размещенной в открытом доступе на веб-ресурсе Drom.ru.

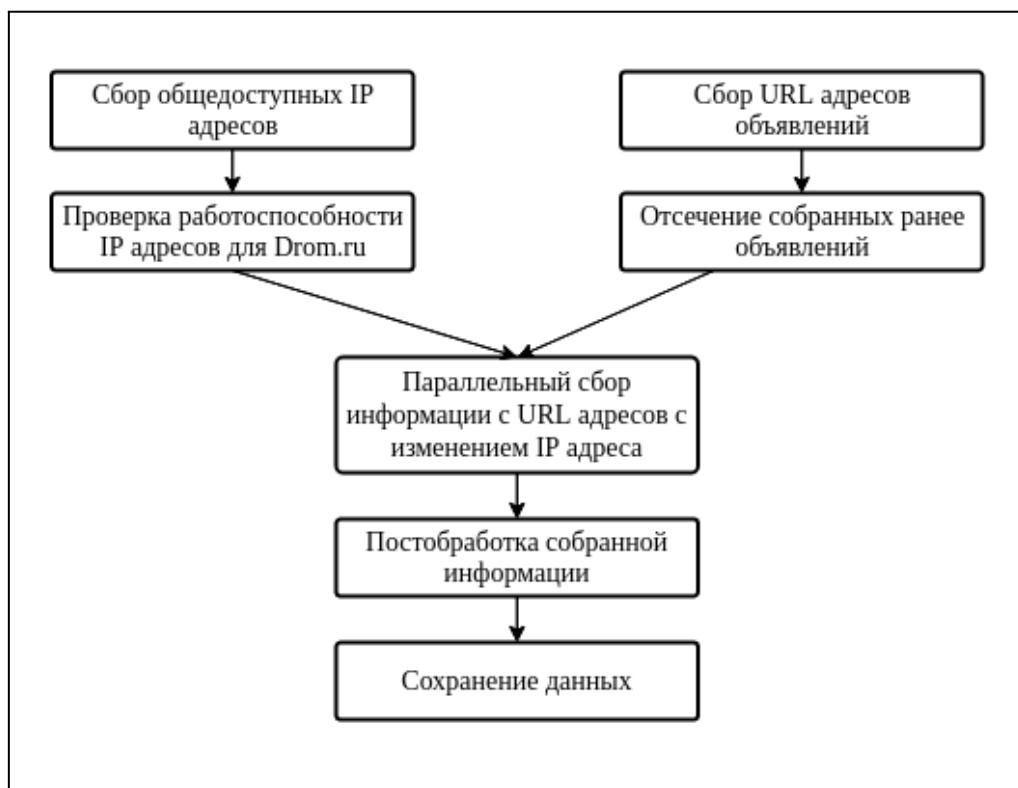
Оценивая снизу время сбора описанной выше информации об объявлениях только автомобилей с сайта Drom.ru, можно сказать, что при затратах в хотя бы в одну секунду на каждое объявление, потребуется более 140 часов непрерывной работы, что является недопустимым, так как данные необходимо собирать на ежедневной основе, не допуская потерю информации об объявлениях самых популярных моделей машин с высоким спросом, которые открываются и закрываются в суточном промежутке. Поэтому необходимо иметь алгоритм, способный обрабатывать не менее 6 страниц веб-ресурса в секунду, чтобы уложиться хотя бы в 24 часа. Имея ограниченные компьютерные мощности и пользовательскую скорость интернет соединения, решено было использовать асинхронный параллельный алгоритм, который может отправлять множество запросов на различные страницы с объявлениями и обрабатывать информацию с них одновременно.

Вместе с этим, стоит учитывать тот факт, что многочисленные частые запросы, отправляемые на сайт, нагружают инфраструктуру самого веб-ресурса, а потому, являются для него нежелательными, вследствие чего сайт блокирует IP адрес пользователя на определенное время, не отвечая на отправляемые сайту запросы. В результате чего, необходимо отправлять ограниченное количество запросов в секунду и менять IP адрес отправителя запросов после очередного определенного количества объявлений на устойчивые адреса общедоступных веб агрегаторов, которые обновляются достаточно часто из-за утраты работоспособности самих IP

адресов. Поэтому, необходимо было сначала создать алгоритм, который может достаточно быстро обновлять список подходящих для сбора информации с сайта Drom.ru IP адресов.

Так как объявления сайта сгруппированы по категориям, например, по бренду, модели, региону. То стоит сначала, изменения данные параметры собирать все URL действующих объявлений, а лишь потом собирать информацию с данных объявлений, делая таким образом срез всех объявлений на конкретный момент времени.

Для того, что выполнять сбор информации с объявлений эффективным образом, необходимо учитывать объявления, которые собирались ранее, а также записывать URL адресов объявлений, которые при отправке запроса получили ошибку соединения, чтобы отправить запрос через некоторое время еще раз.



*Источник: Составлено автором.*

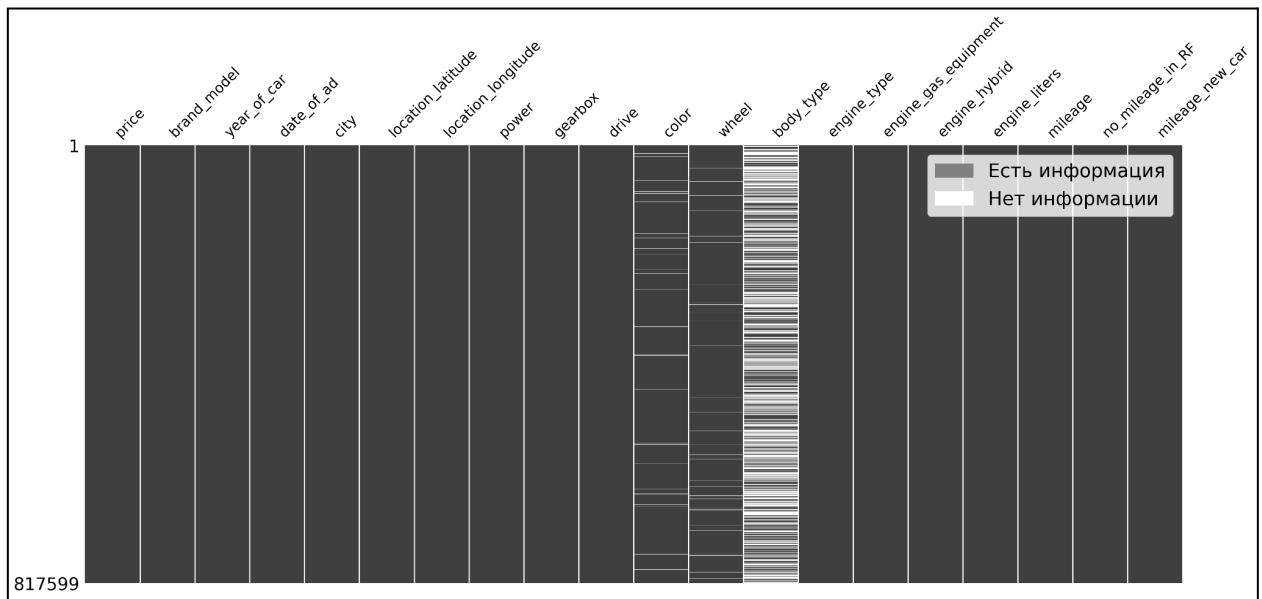
Рисунок 1 – Упрощенный алгоритм сбора и обработки  
объявлений сайта Drom.ru

Собранные данные также необходимо было обработать, так как данные собирались из CSS классов объявления непосредственно. Весомая часть времени здесь была определена под работу с местоположением, так, из прописанной в объявлении локации удалось достать точное адресное местоположение точки продажи и его географические координаты в формате долготы и

широты, что потом позволило соотнести каждое объявление к конкретному региону, а также выполнить географическую кластеризацию.

В итоге, процесс ежедневного сбора и обработки собранной информации в упрощенном формате выглядит следующим образом (рис. 1).

В результате работы парсера за временной промежуток с 2022-10-24 по 2022-12-31 было собрано 987440 уникальных объявлений. После фильтрации данных, а именно, избавления от наблюдений, в которых были пропуски по объему двигателя, мощности, коробке передач, приводу или не было каких-либо данных о пробеге автомобиля, осталось 817599 уникальных объявления.



Источник: Составлено автором.

Рисунок 2 – График полноты наблюдений выборки,  
составленной из объявлений сайта Drom.ru

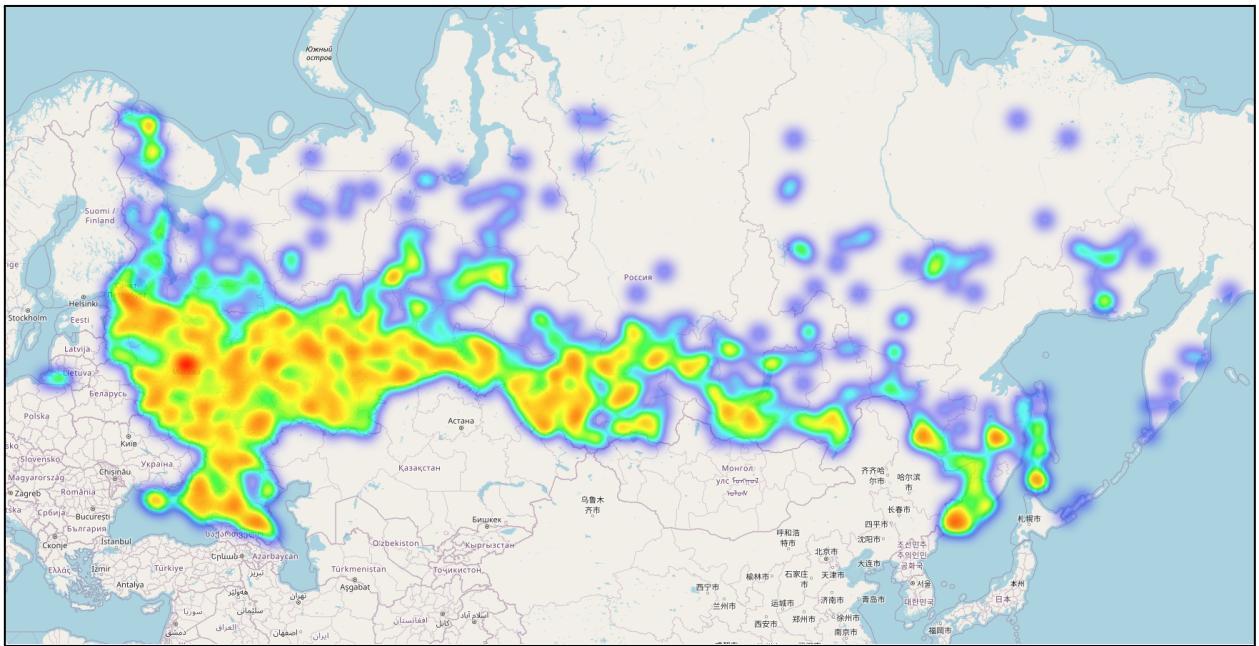
Данная иллюстрация может свидетельствовать о том, насколько собранные данные являются полными относительно каждого наблюдения (рис. 2).

Видно, что количество объявлений, в которых был указан пробег автомобиля достаточно велико, а меньше всего полнота наблюдений в указании типа кузова.

В текущей выборке представлено 3253 уникальных локаций, 124 бренда и 1955 моделей. Однако, после удаления моделей, количество которых в выборке составляет число, меньшее 100, получаем выборку в 790905 наблюдений, где лишь 63 уникальных бренда и 732 уникальных модели. Количество же локаций при этом практически не меняется, теперь число уникальных локаций составляет 3217.

Здесь, можно применить тот факт, что имеется модель каждой машины, а учитывая то, что внутри одной модели тип кузова и сторона руля всегда совпадают, и заполнить пропуски по ним внутри своей группы по модели. В результате, удалось заполнить абсолютно все пропуски по типу кузова и рулевой стороне.

Распределение объявлений по географическим координатам можно представить следующим образом в формате тепловой карты (рис. 3).



*Источник: Составлено автором.*

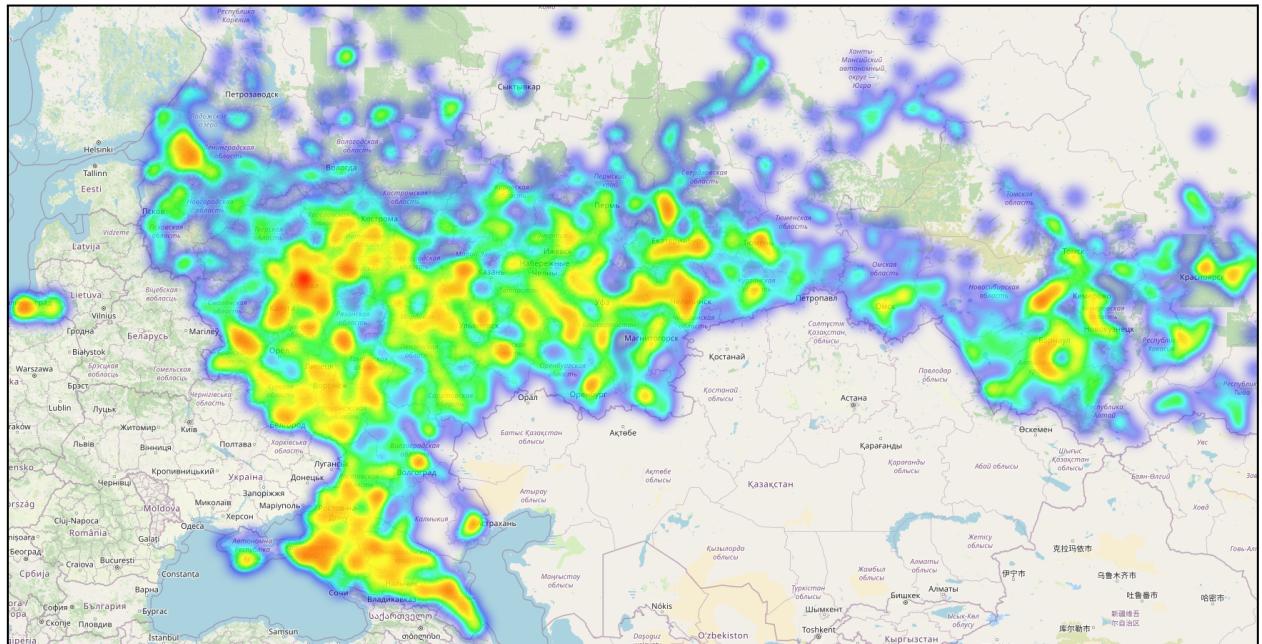
Рисунок 3 – Тепловая карта распределения объявлений автомобилей с сайта Drom.ru  
по географическим координатам

Видно, что географическое распределение собранных данных действительно охватывает практически все населенные пункты Российской Федерации, в том числе и Республику Крым, Калининградскую область, а также соответствует пониманию относительно реально распределенного населения РФ.

Если сосредоточить внимание лишь на наиболее населенных регионах России, то можно увидеть следующую картину (рис. 4). Или же в разрезе центров продаж (рис. 5). В нашей выборке представлено 85 регионов. Наибольшее количество объявлений в нашей выборке приходится на Москву – 69105, Приморский край – 64377 и Новосибирск – 50990, наименьшее количество на Карачаево-Черкесию, Севастополь, Калмыкию, у каждого не более 200 объявлений, а самое малое количество на Ненецкий автономный округ – 3 объявления.

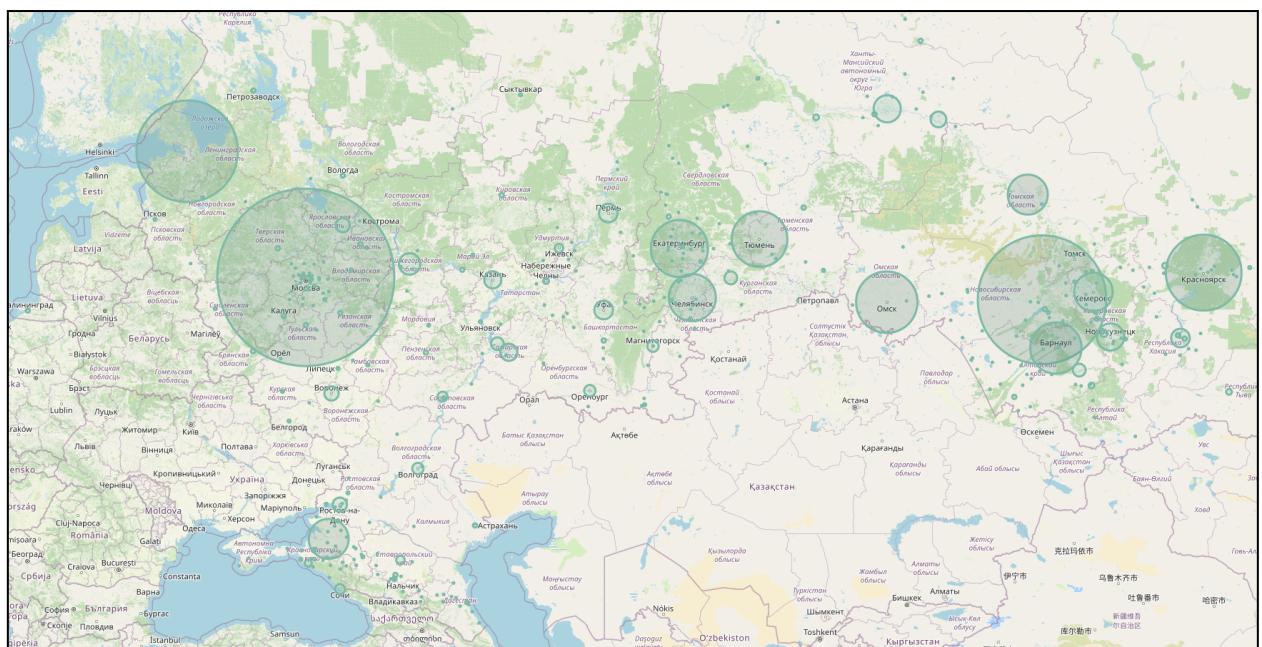
Так как в дальнейшем планируется учитывать индивидуальные эффекты региона, которые не присутствуют в наших региональных данных, необходимо иметь уверенность, что

даные индивидуальные эффекты будут асимптотически близкими и не вводящими в заблуждение. Поэтому, было решено убрать регионы, которые имеют менее 1000 объявлений, всего таких оказалось 21 такой регион. В результате изменения количества регионов с 85 до 64 размер нашей выборки сократился со 790905 до 780365 наблюдений.



*Источник: Составлено автором.*

Рисунок 4 – Тепловая карта распределения объявлений автомобилей с сайта Drom.ru по географическим координатам, западная часть России

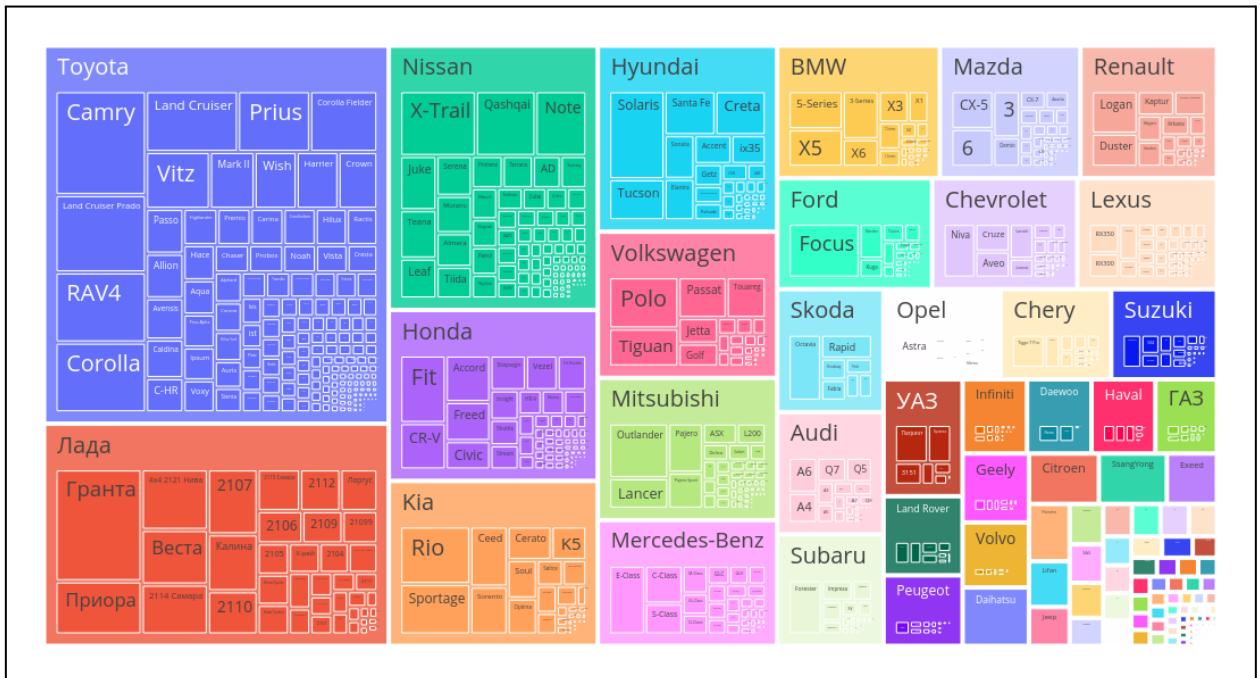


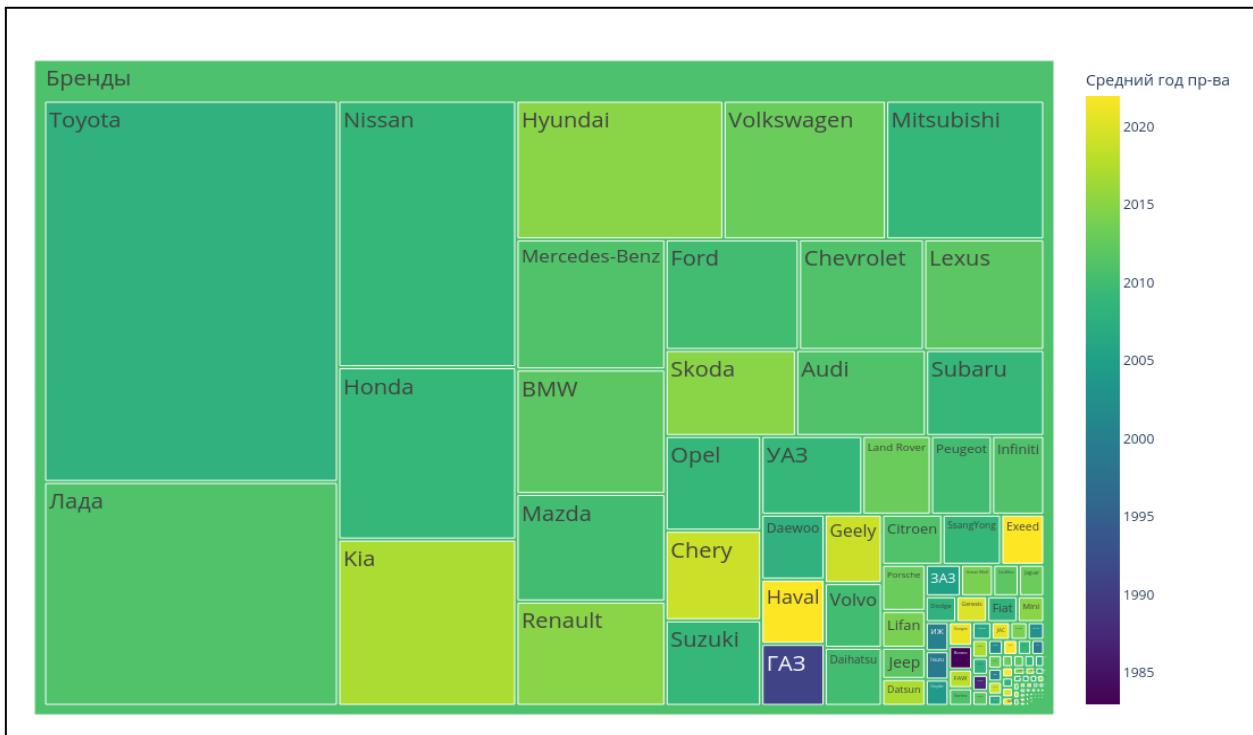
*Источник: Составлено автором.*

Рисунок 5 – Круговая карта центров продаж автомобилей с сайта Drom.ru

по географическим координатам, западная часть России

Так же, посмотрим на распределение представленных брендов и моделей относительно доли наблюдений в текущей выборке (рис. 6). И отдельно посмотрим на распределение брендов по доле наблюдений с учетом среднего возраста автомобиля внутри бренда (рис. 7).

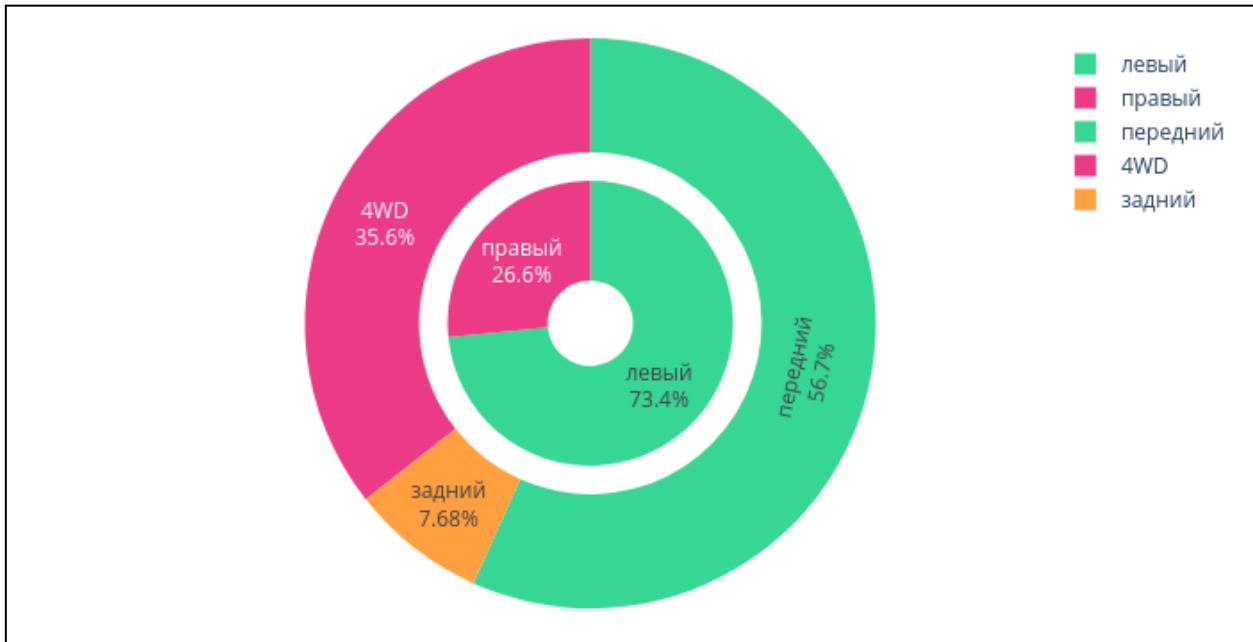




Источник: Составлено автором.

Рисунок 7 – Визуализации комбинированного доли бренда и среднего возраста автомобилей, представленных в нем на сайте Drom.ru

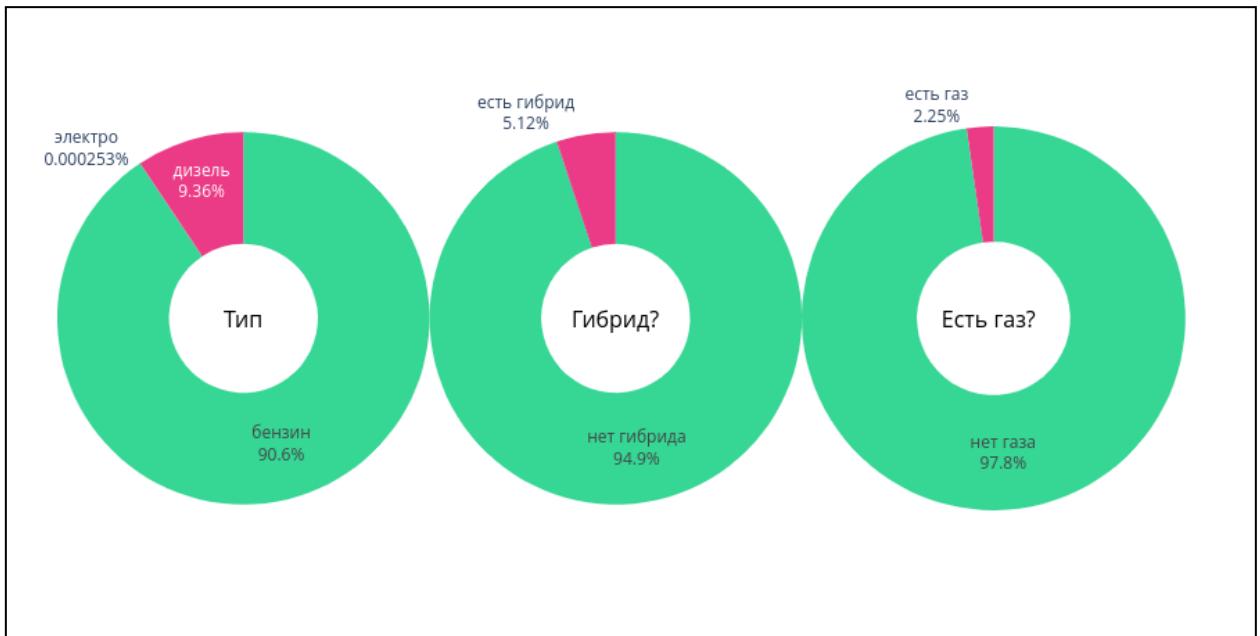
Видно, что наиболее популярными автомобильными марками на площадке Drom.ru являются следующие: Toyota, Лада, Nissan, Honda и Kia.



Источник: Составлено автором.

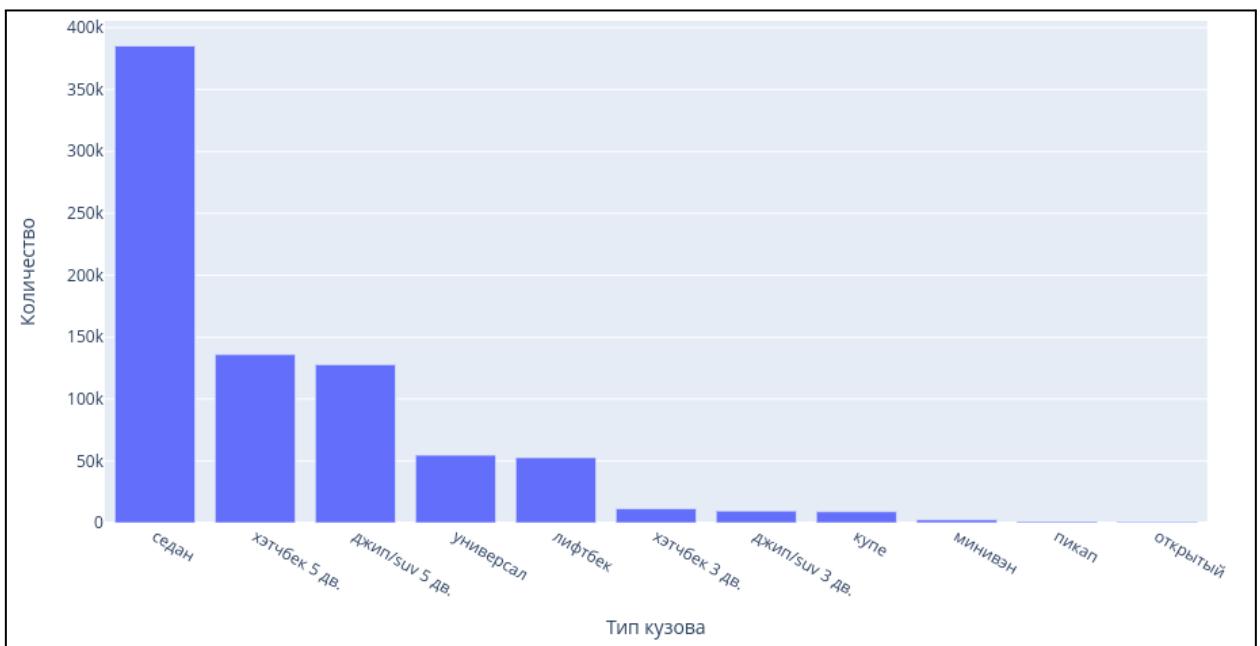
Рисунок 8 – Распределение автомобилей по характеристикам руля и типу привода

Уделим внимание параметру типа кузова, распределение на уже очищенных данных выглядит следующим образом (рис. 10).



Источник: Составлено автором.

Рисунок 9 – Распределение автомобилей по характеристикам двигателя

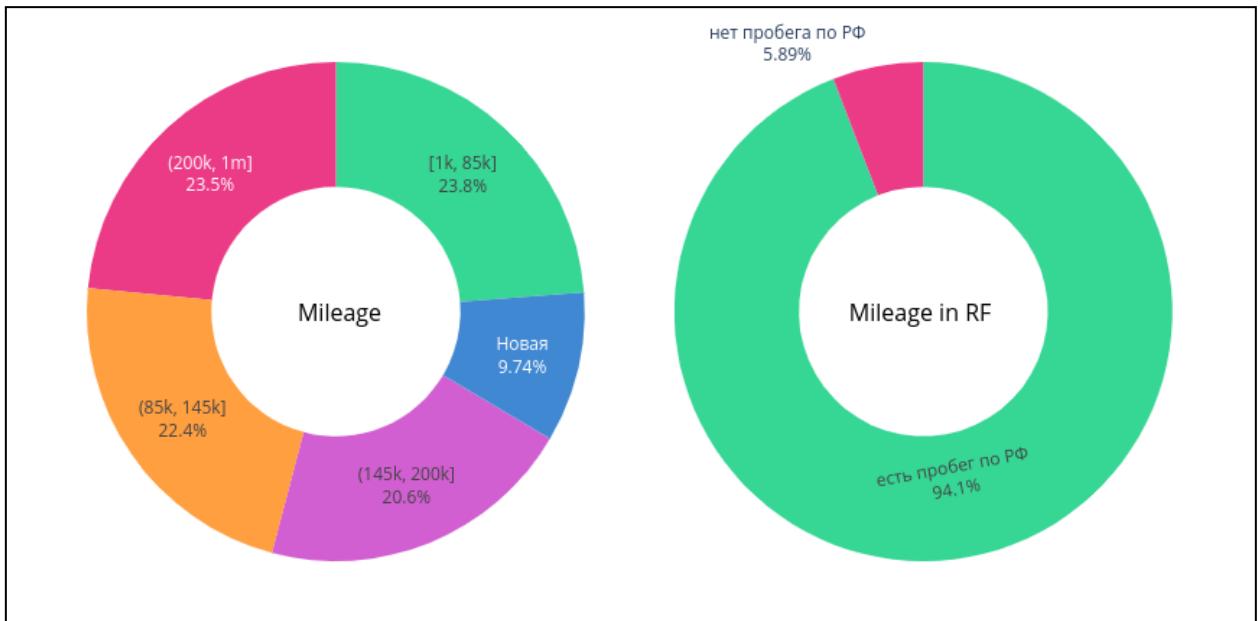


Источник: Составлено автором.

Рисунок 10 – Распределение автомобилей по типу кузова

Видно, что имеется несколько типов кузова с видимым малым количеством наблюдений, а именно: минивэн – 2485, пикап – 1003, открытый – 832. Избавимся от данных типов кузова,

для того чтобы сделать нашу выборку более однородной, в результате, объем выборки сократился до 776098.



Источник: Составлено автором.

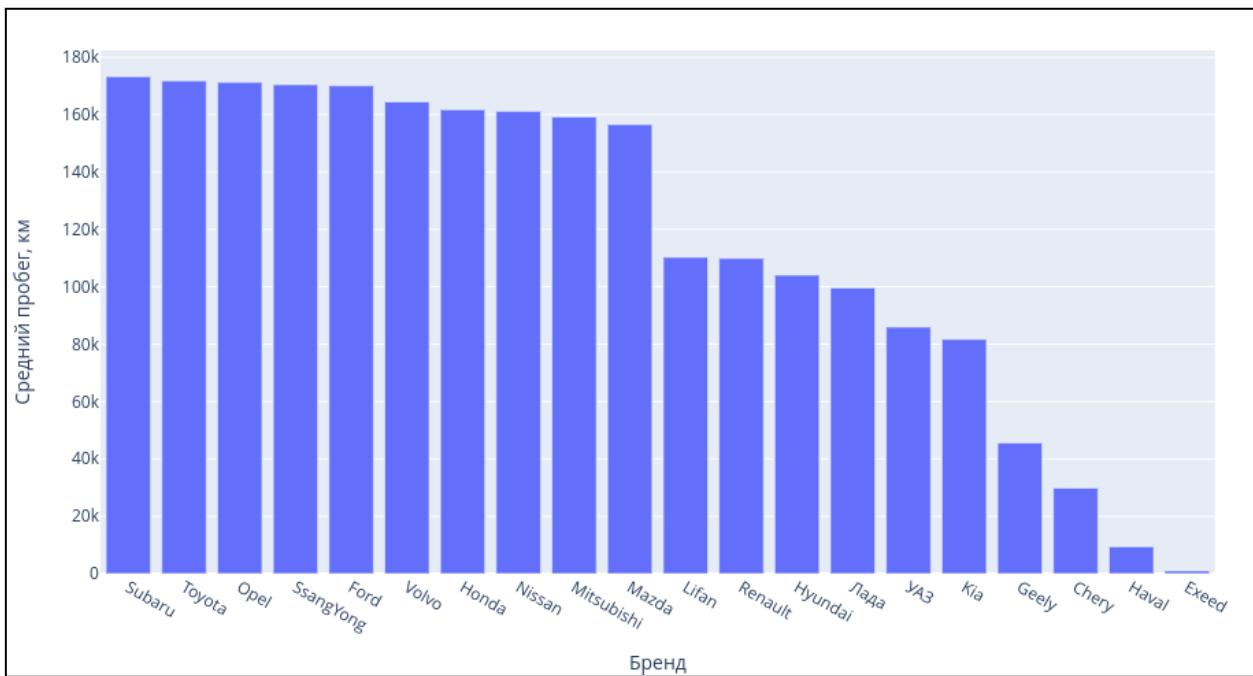
Рисунок 11 – Распределение автомобиля по характеристикам пробега

Также, обратим внимание на очень важный для настоящей квалификационной работы показатель - пробег автомобиля (рис. 11). Как видно, в нашей выборке представлены как новые автомобили, порядка 10%, так и автомобили, которые проехали более 200 тысяч километров. Стоит уточнить, что максимальной величиной пробега, которую можно указать на Drom.ru выступает величина в 999 тысяч километров.

Распределение среднего пробега внутри бренда выглядит следующим образом (рис. 12). Представлены первые 10 и последние 10 брендов по среднему пробегу на выставленных объявлениях на площадке Drom.ru, среди брендов, наличие которых на площадках России не является практически вырожденным, как, например, у Saturn, Xin, Saab, TATA, NiO, Li, объявления которых имеются в выборке, но являются практически единичными.

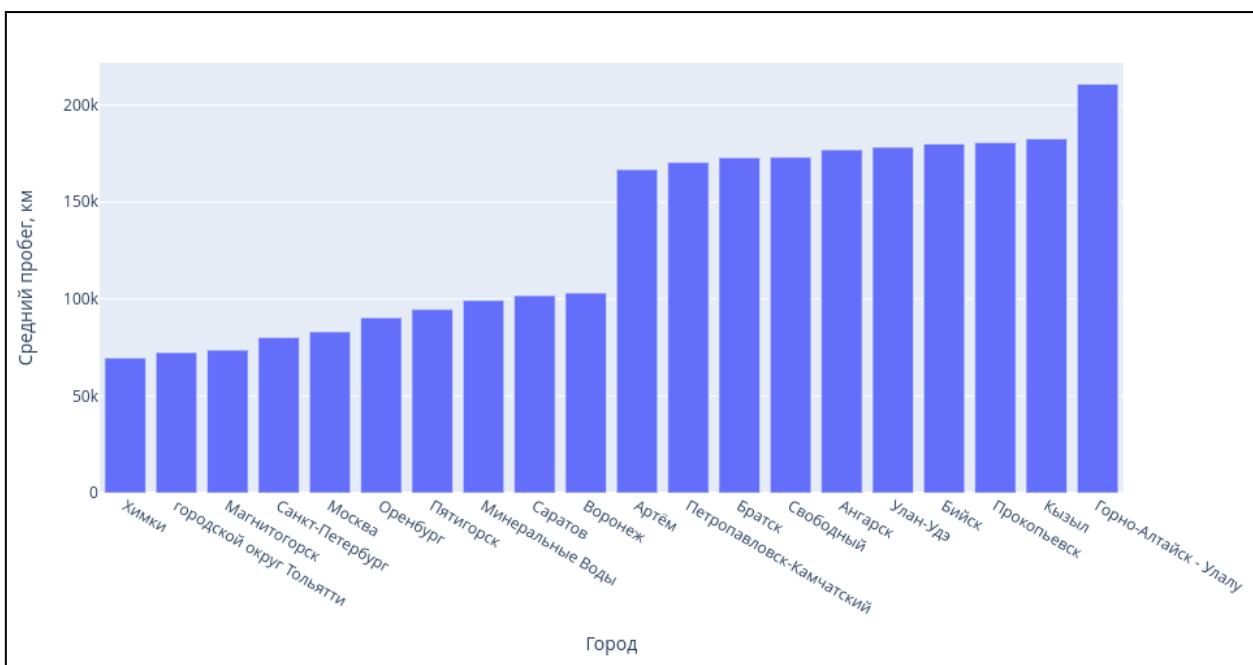
Как видно, наибольший средний пробег встречается среди автомобильных марок, которые имеют наибольшее распространение среди основных масс населения, Лада и УАЗ занимают места с достаточно низким средним пробегом. Первенство по достоинству занимает новый китайский игрок на рынке России – Exceed.

Такой же формат можно увидеть среди первых 10 и последних 10 городов по среднему пробегу (рис. 13).



Источник: Составлено автором.

Рисунок 12 – Первые и последние 10 брендов по среднему пробегу автомобиля



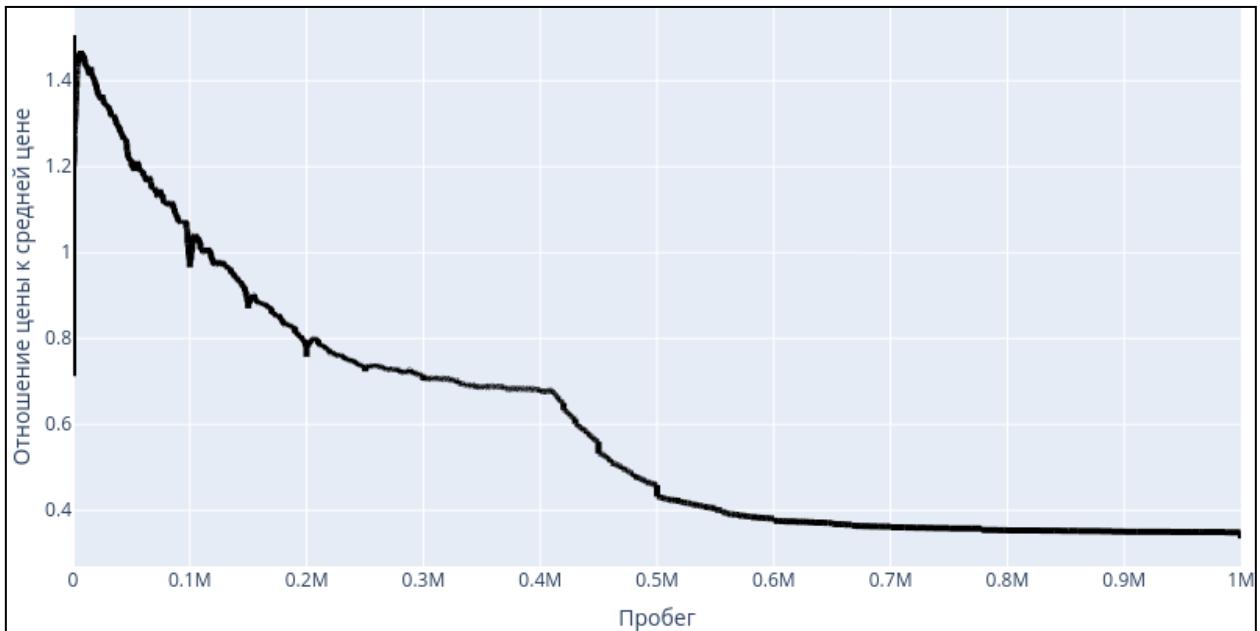
Источник: Составлено автором.

Рисунок 13 – Первые и последние 10 городов по среднему пробегу автомобиля

Такое распределение довольно легко объяснить некоторыми фактами. Первые места с наименее низким средним пробегом занимают либо города-производители автомобилей, то есть, города, в которых находятся крупные заводы по сборке автомобилей, либо города с высоким средним и медианным доходом населения.

В свою очередь, последние города по среднему пробегу на Drom.ru являются отдаленные локации с невысокими доходами населения и местностью, которая подразумевает наличие транспорта с высокой проходимостью.

Уделим большое внимание графику отношения цены к средней цене в скользящем виде с окном в 10 тыс. объявлений, отсортированных по пробегу (рис. 14). Данный график очень важен, поскольку показывает нам то, как в среднем изменяется цена в выборке при изменении пробега.



Источник: Составлено автором.

Рисунок 14 – Скользящая средняя цена автомобиля к средней цене

с окном в 10 тыс. наблюдений по значению пробега автомобиля

Как видно, данное изменение не является линейным, принимая более пологое значение при увеличении пробега автомобиля, наиболее же резкое изменение цены приходится минимальные значения пробега. Интересны также небольшие просадки, приходящиеся на круглые значения – 100 тыс. км., 150 тыс. км., 200 тыс. км., которые можно интерпретировать следующим образом: при отсутствии конкретного значения пробега, который бы владелец автомобиля мог продемонстрировать при продаже и на который бы мог сам опираться при составлении цены в объявлении, продавец склонен занижать выставляемую цену, осознавая, что потенциальный покупатель отнесется более скептически к таким значениям, завышая их.

Также интересен на этом графике и тот факт, что цена автомобиля стабильно изменяется с очень малой скоростью после достижении отметки в 600 тыс. км., что можно отнести на линейную изнашиваемость самих деталей после определенного уровня, где уже роль гарантий и

социального статуса, который может принести автомобиль, практически не влияет на изменение цены при изменении пробега.

Хотелось бы также обратить внимание и на значение в 400 тыс. км., которые выступает порогом для структурного сдвига, что необходимо учитывать при моделировании. То есть, необходимо усечь имеющуюся выборку относительно машин с пробегом, не превосходящим 400 тыс. км., данную необходимость можно также подтвердить тем фактом, что количество автомобилей в нашей выборке после порога в 400 тыс. км. практически затухает и составляет менее 2% от общей доли. Данной операцией нашу выборка уменьшается до размеров в 765883 наблюдений.

В заключении, избавимся от всех наблюдений, которые содержат пропущенные значения, итоговая выборка содержит 750178 наблюдений.

Перейдем к следующему блоку текущего раздела, а именно, к региональным данным, которые будут фигурировать в будущей модели. Так как основной фокус направлен на объяснение цены автомобиля, необходимо иметь подходящие региональные признаки.

В качестве предлагаемых региональных признаков выступают следующие данные социально-экономических показателей конца 2022 года Российской Федерации федеральной службы государственной статистики [Приложение к ежегоднику..., 2023]: население региона, его доля городского населения, протяженность дорог с твердым покрытием на одного человека, количество автобусов на одного человека, средний доход, усредненные по населению кап. вложения за последние пять лет.

Выбран средний доход, а не медианный, так как он показал себя впоследствие лучше при объяснении цены автомобиля, вместе с этим, использование среднего дохода является более предпочтительным по сравнению с медианным, потому что количество приобретаемых автомобилей, а соответственно и количество продающихся на Drom.ru, не ограничивается парадигмой “один человек – один автомобиль”, более обеспеченные индивиды имеют тенденцию к покупке большего количества автомобилей, что могут учитывать средние доходы.

В результате анализа региональных данных было решено прологарифмировать такие показатели, как население региона, средний доход и средние капитальные вложения. Региональные данные были присоединены к имеющейся выборке.

Для того, чтобы эффективно оценить эффект гарантii разных ценовых сегментов, необходимо, во-первых, определить отношение конкретного, имеющегося в выборке, автомобиля к ценовой категории, во-вторых, определить ценообразующее воздействие прочих факторов, чтобы их влияние не учитывалось эффектом гарантii, в-третьих, оценить сам эффект

гарантий для автомобилей внутри ценовой категории с помощью алгоритма, который будет определен в последнем пункте следующей главы.

## **ГЛАВА 2. ВЫБОР ЭКОНОМЕТРИЧЕСКИХ ПОДХОДОВ И МЕТОДОВ И ИХ ТЕОРЕТИЧЕСКАЯ ПОСТАНОВКА**

### **2.1. Модели и методы машинного обучения, применяемые для кластеризации**

Кластерный анализ представляет собой набор исследовательских инструментов, которые могут быть применены при необходимости проверить наличие гомогенного поведения между наблюдениями имеющейся выборки, а также обобщить эти характеристики, создавая определенные правила группировки наблюдений, в которых преобладает внутренняя однородность [Chowdhury, 2017]. В связи с этим основной целью данного набора методов является распределение наблюдений по определенному конечному числу кластеров, которые внутренне однородны и разнородны между собой и которые представляют совместное поведение наблюдений по определенным переменным.

Задача кластеризации относится к задачам обучения без учителя, так как при обучении модели не имеется изначально заданного конечного множества классов, которые присваиваются не обязательно всей, но определенной части выборки. При классификации группы являются априорно зафиксированными, то есть с самого начала имеется понимание, какого рода наблюдения относятся к каждому из них, а также, существует обучающая выборка с примерами объектов и классов, к которым они относятся.

Применимость кластерного анализа в настоящей квалификационной работе будет сведена к осуществлению географической кластеризации объявлений, а также к осуществлению кластеризации ценовых сегментов, которая поможет использовать эконометрические локально в группе наблюдений ценового сегмента. В соответствии с заявленными целями, будет проведена кластеризация сначала по имеющимся географическим координатам, а потом по переменным цены и объема двигателя, которая может служить аппроксимирующей переменной для ценового сегмента.

Инструментарий кластерного анализа может быть отличной заменой методам классификации, когда создание обучающей выборки является слишком дорогой деятельностью или занимает большое количество времени.

Используя методы кластеризации, можно определить оптимальное количество классов для наших данных, если изначально отсутствует предположение, на сколько кластеров должны быть разбиты существующие данные, задать соответствующее распределение классов в имеющейся выборке, а также присвоить классам определенные смысловые обозначения,

которые согласуются с характеристиками самих классов, если их количество умеренное и такая мануальная работа может быть проделана без особых временных затрат.

Простейшими алгоритмами кластеризации являются методы кластеризации с помощью графов, с выделением компонент связности или с составлением минимального оственного дерева. Однако, в силу своей простоты они имеют определенные недостатки, которые не позволяют использовать их в настоящей работе. Среди наиболее популярных различающихся методов кластеризации являются следующие: K-means, Hierarchical Clustering и Density-based spatial clustering of applications with noise, чаще именуемый, как DBSCAN.

Naive K-means, он же алгоритм Лойда [Lloyd, 1982], начинается с априорно задаваемого количества кластеров,  $K$ , то есть, произвольных центров, обычно выбираемых равномерно из пространства данных случайным образом. Затем каждой точке присваивается ближайший центр, объектов выборки распределяются по ним, базируясь на наименьшем евклидовом расстоянии, и каждый центр пересчитывается как центр масс всех назначенных ему точек. Эти два шага, назначение и вычисление центра, повторяются до тех пор, пока процесс не стабилизируется, сойдется. Еще в 2002 году в обзоре методов анализа данных о K-means говорилось, что это “на сегодняшний день самый популярный алгоритм кластеризации, используемый в научных и промышленных приложениях” [Berkhin, 2002]. Данный алгоритм считается наиболее известным методом кластеризации и сейчас.

Hierarchical Clustering, следуя своему названию, является общим классом иерархических алгоритмов, среди самых популярных в котором имеются Agglomerative Hierarchical Clustering и Divisive Hierarchical Clustering. Суть данных алгоритмов легко интерпретируемой, АНС изначально для каждого наблюдения создает собственный класс, через совершение итеративного процесса, имеющиеся классы предыдущей итерации должны быть сгруппированы в меньшее количество классов, базируясь на имеющихся переменных и меры расстояния, данный процесс совершается до тех пор, пока не достигается Критерий останова или не создается конечный класс, объединяющий всю выборку. ДНС наоборот, инициализируется как единый класс для всех наблюдений, который впоследствии разбивается на меньшее количество кластеров, вплоть до одного на каждое наблюдений или ранее, по Критерию останова.

DBSCAN является наиболее продвинутым методом кластеризации, а также, наиболее дорогим, с компьютерационной точки зрения. Данный алгоритм основывается на графовом представлении с выделением связных компонент. Этот алгоритм делит области с достаточной плотностью на кластеры и находит кластеры произвольной формы в пространственной базе

данных с "шумом". Данный алгоритм характеризуется тем, что способен к нахождению сложных кластерных структур, таких как концентрические гиперсфера или протяженные ленты.

Так как цели применения кластерного анализа сводятся к осуществлению выше описанных группировок по цене и географическому признаку, не ожидается увидеть сложные не концентрированные кластерные структуры, поэтому использование DBSCAN было бы излишним для данной работы.

В свою очередь, использование иерархических алгоритмов кластеризации в результате даст несколько кластерных структур, внутри которых может быть осуществлен выбор. За отсутствием необходимости в рассмотрении нескольких вариантов кластеризации, так как преследуется цель лишь найти один оптимальный вариант кластеризации, а также во избежании ситуации, когда НС алгоритм сходится слишком быстро, пропуская эффективный для имеющихся данных вариант кластеризации, было принято решение отказаться от использования НС методов.

K-means является привлекательным для нас вариантом в силу своей простоты интерпретируемости, быстрой сходимости и подходящего формата итоговых результатов работы [Zhao, 2021]. Рассмотрим механизм действия данного алгоритма кластеризации более подробно.

Алгоритм K-means находит К кластеров в соответствии с выбранной метрикой, где обычно используется вариация суммы квадратов ошибок. Временная сложность алгоритма K-means равна  $O(NTK)$ , где N – общее количество наблюдений, K – общее количество кластеров, а T – количество итераций в процессе кластеризации. Эффективность реализации алгоритма K-means относительно высока, но кластеризацию следует выполнять в том случае, если K известно и выбрано, что в большинстве случаев, не так.

K-means минимизирует отклонения внутри кластера, квадраты евклидовых расстояний, но не обычные евклидовы расстояния, что было бы более сложной задачей Вебера: K-means оптимизирует квадраты ошибок, тогда как только геометрическая медиана минимизирует евклидовы расстояния. Например, более удачные по Евклиду решения можно найти, используя K-medians [Jain, 1988] и K-medoids [Kaufman, 1990].

Будет использоваться более усовершенствованный, чем стандартный naive K-means, алгоритм K-means++ из статьи “K-means++: The Advantages of Careful Seeding”, написанной в 2007 году Дэвидом Артуром и Сергеем Вассильвitsким. К сожалению, эмпирическая скорость и простота алгоритма K-means достигается ценой точности. Однако, дополнение к стандартному алгоритму методов K-means++ может помочь как скорости калькуляции K-means, так и результирующей точности.

Итеративная цель алгоритма K-means – уменьшение среднего квадрата расстояния Евклида от центров кластеров до их объектов:

$$\frac{1}{NK} \sum_{k=1}^K \sum_{i=1}^N \rho(\mu_k, x_i) I[a(x_i) = k] \quad (2.1)$$

,  $N$  – количество наблюдений,  $K$  – количество кластеров,  $x_i$  – вектор параметров наблюдения, а  $\mu_k$  – вектор параметров центра  $k$ -го кластера, функция  $a$  – соотносит  $x_i$  ближайшему кластерному центру, функция  $I$  является индикатором, позволяющая оставлять ненулевыми только значения с верно подобранным кластером, базируясь на евклидовом расстоянии, которое рассчитывается следующей общеизвестной формулой:

$$\rho(v_1, v_2) = \|v_1 - v_2\| = \sqrt{\sum_{i=1}^d (v_{1,i} - v_{2,i})^2}. \quad (2.2)$$

При пересчете центры каждого кластера на данном этапе данное условие минимизируется:

$$\sum_{i=1}^N \rho(\mu_k, x_i) I[a(x_i) = k] \quad (2.3)$$

, где  $a(x_i)$  являются фиксированными из шага распределения объектов имеющейся выборки по кластерам.

И при дифференцировании по  $\mu_k$  становится очевидно, что необходимым условием экстремума является переход точки  $\mu_k$  к точке среднего арифметического  $x_i$ , принадлежащих данному кластеру.

Стоит отметить, что, строго говоря, из данных рассуждений не следует, что обязательно глобальный минимум будет найден оптимизирующей функции, так как это, как минимум, зависит от изначального распределения кластеров. Однако, можно с гарантией утверждать, что при заданном начальном распределении центров всегда имеется возможность найти соответствующий локальный минимум, который, конечно, может оказаться далек от глобального, за конечное число итераций, при условии, что K-means не войдет в бесконечный цикл.

K-means++ специфицирует выбор начальных центроидов, отходя от использования примитивной инициализации с использованием равномерного распределения на имеющейся

выборке. Пусть  $D(x_i)$  является наикратчайшим расстоянием от точки  $x_i$  до ближайшего центра кластера. Тогда алгоритм K-means++ может быть описан следующим образом:

1. Случайно, используя равномерное распределение, инициализировать центр первого кластера  $c_1$ , основываясь на подающейся выборке  $X$ ;
2. Выбрать центр следующего кластера  $c_i$ , где  $c_i = \hat{x} \in X$  с вероятностью

$$\frac{D(\hat{x})^2}{\sum_{x \in X} D(x)^2}. \quad (2.4)$$

3. Повторить 2 этап пока все К центроиды не будут инициализированы;
4. Перейти к выполнению шагов naive K-means.

В результате, на эмпирических данных, было выявлено, что такая инициализация центров кластеров в среднем ускоряет время работы алгоритма приблизительно в 10 раз и улучшает метрику приблизительно как минимум на 10% [Arthur, 2007]. Вместе с этим, в работе “K-means++: The Advantages of Careful Seeding” проводится строгое математическое доказательство эффективности приводимых выше операций по сравнению с наивной инициализацией центроидов.

Для того, чтобы использовать алгоритмы кластеризации часто прибегают к методам стандартизации имеющихся величин, для того, чтобы избежать ошибок кластеризации, которые могут быть порождены различной размерностью переменных. Это следствие того, что K-means кластеризация изотропна во всех направлениях пространства и, следовательно, имеет тенденцию к образованию более или менее округлых кластеров, не эллипсоидоподобных. В этой ситуации оставление дисперсий неравными эквивалентно приданию большего веса переменным с меньшей дисперсией.

В нашем случае, будет использоваться общераспространенная формула стандартизации:

$$z = \frac{x - \theta}{\sigma} \quad (2.5)$$

, где

$$\theta = \frac{1}{N} \sum_{i=1}^N x_i, \quad (2.6)$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \theta)^2}. \quad (2.7)$$

Для того, что определить оптимальное количество кластеров, используя алгоритм K-means++, прибегают к различным статистикам. Наиболее распространенной является легко интерпретируемое значение суммы квадратов ошибок внутри кластера, именуемое Interia или WWS, от within-cluster-sum of squares errors. То есть, внутри каждого кластера находится сумма квадратов расстояний от точки до центроида, после чего они суммируются и получается WWS. На основе данной статистики строится так именуемая Elbow curve, которая содержит значения WWS, приходящиеся на конкретное количество кластеров. Истоки использования данного метода можно проследить в ранних работах по психометрии [Thorndike, 1953].

Суть данного графика лежит в нахождении некоторой точки, Elbow point, в которой среднеквадратическое отклонение от центроидов падает наиболее сильно, и после чего изменение WWS начинает затухать, напоминая положение руки в согнутом состоянии. Именно поэтому данный график носит такое название, напоминая “изгиб локтя”.

Данный метод, помимо общей распространенности, нашел и критику, которая признает Elbow curve как субъективный и ненадежный метод [Ketchen, 1996], а выбор "колена" во многих практических приложениях весьма неоднозначен, поскольку на графике нет острого падения статистики, для нахождения Elbow point.

Также, достаточно распространенной статистикой для отбора оптимального количества кластеров является так называемая GAP статистика, которая была предложена в 2001 году в работе “Estimating the number of clusters in a data set via the gap statistic” под авторством Роберта Тибширани, Гюнтера Вальтер, Тревора Хести.

Данный метод, как и Elbow curve, использует выходные данные любого алгоритма кластеризации, сравнивая изменение дисперсии внутри кластера с ожидаемым при соответствующем эталонном нулевом распределении. Для этого была разработана соответствующая теория, и имитационное исследование авторов показывает, что GAP статистика обычно превосходит другие методы, которые были предложены в литературе до этого.

Обнаружение оптимального по GAP количества кластеров не требует определенного рода спекуляций, как того в некоторых ситуациях может потребовать Elbow curve при обнаружении Elbow point, напротив, оптимальным по GAP количеством кластеров считается число, приходящееся на максимальное значение GAP статистики, что сильно упрощает его нахождение.

Следуя вышесказанному, будут сравниваться результаты выбора оптимального количества кластеров по WSS и GAP статистикам. При наличии разнотений среди этих двух статистик, если таковые появятся в данной квалификационной работе, среди данных инструментов будет выбираться то число кластеров, которое имеет более близкую к реальности интерпретацию.

В результате нахождения оптимального количества кластеров строиться заключительная модель K-means++ по стандартизованным величинам, возвращая метки классов для каждого наблюдения, а также инверсивно трансформированные результаты - значения величин центроидов.

Вместе с этим, найденные ценовые сегменты для каждого наблюдения могут образовать разнотения внутри категории бренд-модель, поэтому, чтобы этого избежать и поставить истинное соответствие между моделью и ценовой категорией, нужно присвоить каждой модели модальную ценовую категорию внутри своей группы.

## **2.2. Эконометрические методы пространственного анализа**

В данном разделе будут рассмотрены элементы пространственной эконометрики, которые были использованы в настоящей квалификационной работе.

Начало пространственного анализа состоит в построении пространственной матрицы, основанной на географической близости регионов, которую будем обозначать как  $W$ .

Элементы данной матрицы являются обратными расстояниями между центрами регионов в километрах, которые были получены ранее при формировании данных.

После постройния  $W$ , необходимо провести нормировку по строке, как это негласно принято в эконометрическом сообществе. Сторого говоря, отсутствие нормирования, здесь под нормированием понимается только деление на сумму по строке, никак не влияет на корректность пространственных эконометрических моделей, так как данное действие влияет лишь на масштабируемость конечной оценки для коэффициента при матрице пространственных весов.

После того, как матрица  $W$  оказалась рассчитанной, необходимо построить индикатор пространственной зависимости, в роли которого, выступает наиболее распространенный в эконометрическом сообществе глобальный индекс Морана I [Moran, 1950], где считается, что при значимой величине индекса его положительность, относительно своего математического ожидания, свидетельствует о положительной пространственной корреляции, то есть, регионы считаются окружеными подобными себе по значениям зависимой переменной.

Раскроем механизм данного индекса более подробно. При нулевой гипотезе о том, что данные являются независимыми и одинаково распределенными нормальными случайными величинами, распределение I Морана известно, и тесты гипотез, основанные на этой статистике, обладают различными свойствами оптимальности. Учитывая его простоту, I Морана также часто используется вне рамок формальной проверки гипотез при исследовательском анализе данных с пространственными привязками.

Глобальная статистика Морана может быть не столь удачной при расчете пространственных авторегрессионных моделей, как например, заявляется в работе 2007 года “Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model” под авторством Х. Ли, К. А. Колдер и Н. Кресси, в которой они предлагают использование Approximate profile-likelihood estimator, APLE, и утверждают, что данный тест более оптимален для оценки данных, подходящих для Spatial autoregression, SAR, модели. Однако, в силу распространенности использования статистики Морана, а также по причине отсутствия имплементации APLE в библиотеках Python и сложности в собственном имплементировании, было решено ограничиться использованием I Морана для тестирования наличия пространственной автокорреляции.

Рассчитывается данная статистика Морана следующим образом:

$$I = \frac{\sum_{i=1}^N \sum_{j=1}^N \omega_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i,j} w_{ij} \sum_{i=1}^N (x_i - \bar{x})} \quad (2.8)$$

, где  $N$  – количество пространственных единиц, проиндексированных через  $i$  и  $j$ ,  $x$  – объясняемая переменная для расчета индекса,  $\bar{x}$  – выборочное среднее  $x$ ,  $w_{ij}$  – элементы матрицы  $W$ .

Данная статистика принимает значения в интервале  $[-1, 1]$ , а под условиями нулевой гипотезы математическое ожидание данной величины равняется следующему значению:

$$E(I) = \frac{-1}{N-1} \quad (2.9)$$

И, как уже говорилось ранее, если  $I > E(I)$  при условии 5% значимости данной статистики, говорят о положительной пространственной автокорреляции.

Имея данные о сегментированных по цене объявлениях автомобилей, которые структурированы по региону и модели автомобиля, предлагается использовать Spatial Lag Model,

SLM, которая также именуется Spatial Autoregression model, SAR. Данная модель учитывает пространственный лаг для объясняющей переменной при помощи матрицы  $W$ .

Однако, учитывая особенности имеющихся данных, стоит уточнить, что матрица пространственных весов умножается не на сам вектор объясняющих переменных, так как вне регионального представления данных имеется лишь набор объявлений, не подлежащий векторизации. Матрица  $W$  в настоящем случае умножается на вектор средних внутрисегментных цен автомобилей внутри региона. Где в отдельном случае конкретного объявления данной переменной присваивается значение данного пространственного лага для региона объявления.

Данная модель будет строиться в панельном виде, учитывая фиксированные эффекты, которые оказывает модель автомобиля. Изначальный выбор именно фиксированных эффектов, без рассмотрения возможности использования случайных эффектов обуславливается следующим рассуждением.

Модели со случайными и фиксированными коэффициентами, далее, RE и FE, функционально связаны. Для этого можно посмотреть на оценки данных моделей.

Модель с FE можно записать в стандартном виде следующим образом:

$$y_{ij} = \alpha_i + x_{ij}'\beta + \epsilon_{ij}, \epsilon_{ij} \in IID(0, \sigma_\epsilon^2) \quad (2.10)$$

, где  $i$  соответствует модели автомобиля, а  $j$  номеру наблюдения для данной модели,  $\alpha_i$  является оценкой для параметра фиксированного эффекта от  $i$ -й модели.

Также нам понадобится для понимания эквивалентная запись данной регрессии в следующей форме:

$$y_{ij} - \bar{y}_i = (x_{ij} - \bar{x}_i)' \beta + (\epsilon_{ij} - \bar{\epsilon}_i) \quad (2.11)$$

, где значения с чертой соответствуют выборочным средним по  $j$ , то есть, по всем наблюдениям конкретной модели.

В результате, оценка коэффициентов для регрессии с фиксированными эффектами, так же именуемая Least Square Dummy Variable, LSDM, или Within Estimator, WE, выглядит следующим образом:

$$\hat{\beta}_{FE} = \left( \sum_{i=1}^N \sum_{j=1}^M (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \right)^{-1} \sum_{i=1}^N \sum_{j=1}^M (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i) \quad (2.12)$$

, а оценки для самих фиксированных эффектов можно получить следующим образом:

$$\widehat{\alpha}_i = \bar{y}_i - \bar{x}_i' \widehat{\beta}_{FE} . \quad (2.13)$$

Учитывая слабые условия регулярности, FE оценки считаются асимптотически нормальными, в результате чего могут быть использованы стандартные вариации теста Вальда, например, тест на значимость регрессора, использующий t-статистику.

Оценка коэффициента для RE модели, которая является General Least Square оценкой [Maddala, 1971], является средневзвешенным так называемой Between Estimator, для регрессии, усредненной по времени, и FE. Где матрица весов определяется как пропорционально обратная к ковариационной матрице Between Estimator.

Вместе с этим, связь между FE и RE оценками уменьшается при стремлении  $J$  к бесконечности и фиксированном  $N$ . Поэтому, для выборок с большим  $J$  различия между двумя этими моделями пропорционально уменьшаются. А для выборок с малым  $J$  и достаточно большим  $N$  различия между результатами этих двух моделей могут приобретать значимый характер, как например, Хаусман обнаружил [Hausman, 1978], что использование спецификации с фиксированными эффектами дает значительно отличающиеся результаты от спецификации со случайными эффектами при расчете уравнения заработной платы с использованием выборки из 629 выпускников средней школы, за которыми в течение шести лет велось мичиганское исследование динамики доходов.

Теперь обратимся к интерпретационной и смысловой составляющей двух вышеописанных моделей. Основное различие моделей состоит в том, что в модели FE распределение объясняемой переменной является условным по значениям эффектов. Интуитивно такая интерпретация имеет смысл, если индивидуумы, фирмы или регионы, в выборке являются близкими к единственным в своем роде, и не могут рассматриваться как случайные извлечения из некоторой лежащей в основе генеральной совокупности, а соответственно и быть экстраполированы за нее.

Напротив, подход RE не является условным по индивидуальным эффектам, и в этом случае обычно нет заинтересованности в конкретном значении эффекта для некоторого индивидуума, но есть сфокусированность на случайно выбранных, которые имеют определенные характеристики.

Применяя приведенную выше теорию, можно сделать вывод о том, что модель RE не стоит рассматривать по наличию того факта, что в имеющейся выборке рассматриваются уникальные модели различающихся брендов автомобилей.

В результате, наша модель будет компоноваться из SLM на панельных данных с фиксированными эффектами по модели автомобиля, количество которых равняется количеству ценовых сегментов.

Функциональная запись данной модели выглядит следующим образом:

$$\begin{aligned} \log(price_{ij}) = & \alpha_i + state_{ij}'\beta_{state} + car\_cont_{ij}'\beta_{cont} + \\ & + car\_dummies'\beta_{dummies} + \varrho w_{ij} + \epsilon_{ij}. \end{aligned} \quad (2.14)$$

Которая будет оценена при помощи эквивалентной формулы с вычислением средних значений объясняющих, кроме категориальных, и объясняемой переменных на промежутках принадлежности наблюдения к конкретной автомобильной модели, то есть, в следующем виде:

$$\begin{aligned} (\log(price_{ij}) - \overline{\log(price)})_i = & (state_{ij} - \overline{state}_i)' \beta_{state} + \\ & + (car\_cont_{ij} - \overline{car\_cont}_i)' \beta_{cont} + car\_dummies'\beta_{dummies} + \varrho(w_{ij} - \overline{w}_i) + \epsilon_{ij} \end{aligned} \quad (2.15)$$

Переменные, которые будут использованы в вышеприведенной записи модели, будут следующие вещественные переменные: значение вектора-матрицы  $W$  для региона, обозначенные ранее переменные региона, мощность двигателя, его объем, пробег, а также корень пробега, который удачно описывает кривую скользящего среднего цены по пробегу (рис. 14). А также категориальные переменные с базовой фиктивной переменной самой распространенной категории: типа коробки передач – АКПП, привода – передний, рулевой стороны – левый, тип кузова – седан, цвет автомобиля – белый, типа двигателя – бензин, а также фиктивные переменные наличия газового оборудования и переменная гибридности двигателя.

### 2.3. Алгоритм определения эффекта гарантii

В данном пункте будет представлен алгоритм, в соответствии с которым оценивались эффекты гарантii для автомобилей разных ценовых сегментов.

Сначала определим то, как стоит оценивать эффект гарантii. Эффект гарантii в области торговли представляет собой величину, на которую, или в которую, цена товара или услуги была больше, при условиях, что субъект, осуществляющий предложение данного блага, обладает

инструментами воздействия на покупателя, посредством которых продавец, чтобы заручиться некоторым доверием покупателя, берет на себя ответственность за них.

В нашем случае, лицом, предоставляющим гарантиями может выступать автомобильный дилер, предоставляющий предложение новых и подержанных автомобилей, или официальные комиссионные автоцентры, которые осуществляют покупку, возможный ремонт, и продажу автомобилей.

В связи с особенностью анализируемых данных, в которых присутствуют объявления о продаже новых автомобилей без пробега, мы можем оценить эффект гарантiiй, если нам удастся посчитать так называемый Average Treatment Effect, ATE, который также именуется как Average Causal Effect или ACE, в соответствии с общей моделью RCM, которая будет описана ниже. В случае оценки эффекта гарантiiй, контрольной группой будут выступать автомобили, которые имеют пробег, а потому, объявления которых выставлены физическими лицами, а группой воздействия автомобили, которые не имеют пробега и в подавляющем количестве выставляются официальными юридическими лицами, которые способны предоставить гарантiiи.

Одним из основных предположений для осуществления оценки какого-либо эффекта является так называемое предположение стабильности величины воздействия, Stable Unit Treatment Value Assumption, которое утверждает, что воздействие имеет эффект только на группу воздействия и не имеет никаких, в том числе непрямых, эффектов на контрольную группу.

Поэтому, для того, чтобы правильно определить эффект гарантiiй между данными двумя группами, необходимо нивелировать эффект, который оказывается пробегом автомобиля на стоимость самого транспортного средства, и сравнивать средние очищенные от пробега цены внутрисегментных автомобилей. Для этого, будут использоваться параметры из панельной SLM с фиксированными эффектами, в которой оценивается воздействие, оказываемое пробегом на цену автомобиля. Вместе с этим, для того, чтобы сделать группы более однородными, при этом, не слишком сильно сокращая имеющиеся для оценки эффекта гарантiiй объем данных, необходимо использовать для контрольной группы лишь автомобили, которые прошли менее 10 тыс. км., и не получили высокого прямого износа.

В то время как эксперимент гарантирует, что потенциальные исходы будут эквивалентно распределены в группах воздействия и контрольной группе, в наблюдаемом исследовании, коим настоящая квалификационная работа является, этого не происходит. В эмпирическом исследовании отдельные субъекты выборки не назначаются в вышеупомянутые группы случайнym образом, поэтому их назначение может зависеть от ненаблюденных факторов. Данное предположение, необходимое для осуществления оценки эффектов, можно встретить в

русскоязычной эконометрической литературе под названием общности оснований, Common Support.

Для того, чтобы избежать данную проблему, часто прибегают к методам мэтчинга, которые позволяют получить гомогенные подгруппы, из имеющихся в выборке групп и посчитать Sample ATE, SATE, с помощью которых можно взвешенно по размеру подгрупп рассчитать ATE.

Первые идеи мэтчинга в эконометрическом сообществе появились в работе Дональда Б. Рубина 1973 года “Matching to Remove Bias in Observational Studies”, которые впоследствии развивались многими ведущими эконометристами [Anderson, 1980; Kupper, 1981] и развиваются до сих пор [King, 2019].

Репрезентативные примеры использования методов мэтчинга продемонстрированы в широко известных работах “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme” [Heckman, 1997] и “Propensity Score-Matching Methods for Nonexperimental Causal Studies” [Dehejia, 2002].

Наиболее эффективным методом мэтчинга является так называемый точный мэтчинг, Exact Matching, в соответствии с которым, имеющиеся группы воздействия и контроля делятся на подгруппы по признаку точного совпадения значений по выбранным исследователем переменным. В случае настоящей работы, хорошим параметром для точного совпадения является модель автомобиля.

Однако, внутри одной модели автомобиля, могут находиться объекты разной комплектации с различающимися характеристиками, как собственными, так и не собственными – характеристиками региона. Для того, чтобы учесть данные особенности составления подгрупп, может применяться следующий метод мэтчинга.

Мэтчинг по индексу склонности, Propensity Score Matching или PSM, является распространенным в эконометрическом сообществе методом мэтчинга, механизм действия которого был исследован в работе 1983 года “The Central Role of the Propensity Score in Observational Studies for Causal Effects” под авторством П. Р. Розенбаума и Б. Р. Рубина.

Данный подход являются частью общей, так называемой Rubin Causal Model, RCM, также известной, как Neyman-Rubin Causal Model [Sekhon, 2009], которая, вместе с теоретической постановкой мэтчинга по индексу склонности, имела критику в уже упомянутой недавней статье Г. Кинга и Р. Нильсена 2019 года “Why Propensity Scores Should Not Be Used for Matching”, в которой предлагается простой, понятный и статистически мощный метод сопоставления, известный как Coarsened Exact Matching или CEM. RCM является подходом к статистическому

анализу причин и следствий, основанный на структуре потенциальных результатов, описываемых через вышеупомянутые внутригрупповые эффекты.

При использовании PSM существует несколько методов разбиения контрольной группы и группы воздействия на подгруппы на основе индекса склонности, среди них наиболее популярные следующие: Nearest Neighbor Matching, Stratification Matching, Difference-In-Differences Matching. Будет использоваться Stratification Matching, который предполагает разбиение на несколько страт, которые были названы выше подгруппами, внутри которых все объекты обладают примерно одинаковой вероятностью подвергнуться воздействию.

Использование нескольких методов мэтчинга, в настоящем случае – точного мэтчинга и мэтчинга по индексу склонности со стратификацией, часто обобщенно именуется смешанным мэтчингом.

Общий применяемый в настоящей работе механизм мэтчинга выглядит следующим образом.

На первом этапе производиться точный мэтчинг по модели автомобиля, в результате которого образуются определенные блоки объявлений автомобилей, которые принадлежат контрольной группе и группе воздействия.

Первым шагом в алгоритме PSM является построение модели бинарного выбора для присвоения индекса склонности каждому наблюдению, в эконометрическом сообществе принято использовать регрессионные модели бинарного выбора, а именно логит или пробит, результат присвоения которых слабо отличается между собой. Для настоящей квалификационной работы была выбрана логистическая модель, в силу своей большей распространенности.

Признаками для модели логита выступили описанные выше переменные, характеризующие регион местонахождения наблюдения, значение пространственного лага для региона, а также переменные самого автомобиля кроме тех, которые характеризуют его изношенность – коробка передач, привод, тип двигателя и его объем, мощность, руль.

В результате первого шага PSM мы получаем внутри каждого блока по модели автомобиля индекс склонности для каждого наблюдения. Функциональная запись приведенной бинарной логит модели выглядит следующим образом:

$$\text{Propensity Score} = P(y_i = 1|X) = \frac{1}{1+e^{-X\beta}}, \quad (2.16)$$

где  $y_i = \{0, 1\}$  – факт принадлежности автомобиля к группе воздействия, то есть, факт того, что автомобиль – новый, а  $X$  – матрица, содержащая вышенназванные переменные. Оценивается данная модель с помощью метода максимального правдоподобия.

Далее, формируется контрольная группа из индивидов, не подвергшихся воздействию, с наиболее близкими значениями индекса склонности для наблюдений из группы воздействия.

В результате данного смешанного мэтчинга находятся страты наблюдений с группами воздействия и контрольными группами, используя которые, можно рассчитывать SATE, а впоследствии, и ATЕ для каждого ценового сегмента.

## ГЛАВА 3. ОПИСАНИЕ И ОЦЕНКА ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

### 3.1. Краткая характеристика кластерного анализа

Изначальное назначение гиперпараметров, необходимых для инициализации алгоритма K-means++ было выбрано в соответствии с пониманием имеющихся данных. Так, максимальное количество итераций, которое было установлено для алгоритма, равнялось 300, гиперпараметр количества повторений является заданным и соответствует единице, по причине выбора алгоритма K-means++, дающего детерминистический результатов первоначального выбора центроидов, по причине чего получаемый результат распределения наблюдений по кластерам не нужно повторять на сходимость к нему дополнительными повторениями.

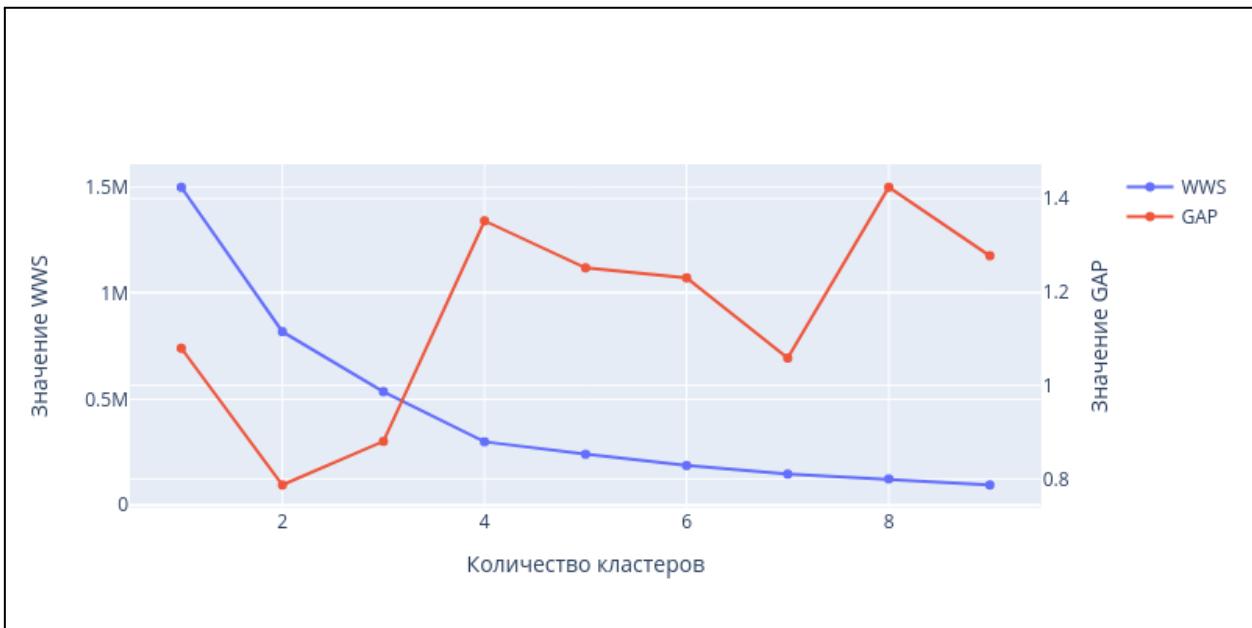
Для того, чтобы найти оптимальное количество кластеров, используя Elbow Curve и GAP, необходимо иметь результаты распределения для всех чисел кластеров, а соответственно, они должны быть построены.

Исходя из эмпирических представлений об имеющихся данных, максимальным количеством для пространственного распределения было выбрано величина в 9 кластеров, а для распределения по ценовым сегментам в 7 кластеров.

Описанные во второй главе настоящей работы Elbow Curve, оптимальной точкой для которого является так называемая Elbow point, и GAP, с оптимальным количеством кластеров при наибольшем значении величины GAP, выглядят следующим образом для пространственной кластеризации (рис. 15).

Вместе с этим, в используемый в данной работе программный инструментарий имплементирован метод, который позволяет численно вычислить Elbow Point для подающихся значений, в текущем случае, данное значение равно 4, что действительно соответствует визуальной точке “сгиба локтя”.

Так же видно, что значение GAP принимает локальный максимум для 4 кластеров, что говорит о потенциально подходящем выборе, однако, максимально значение на всем промежутке достигается на значении в 8 кластеров. Сообразуясь с результатами данных методов, стоит выбрать определенный компромисс в количество кластеров, равное четырем.

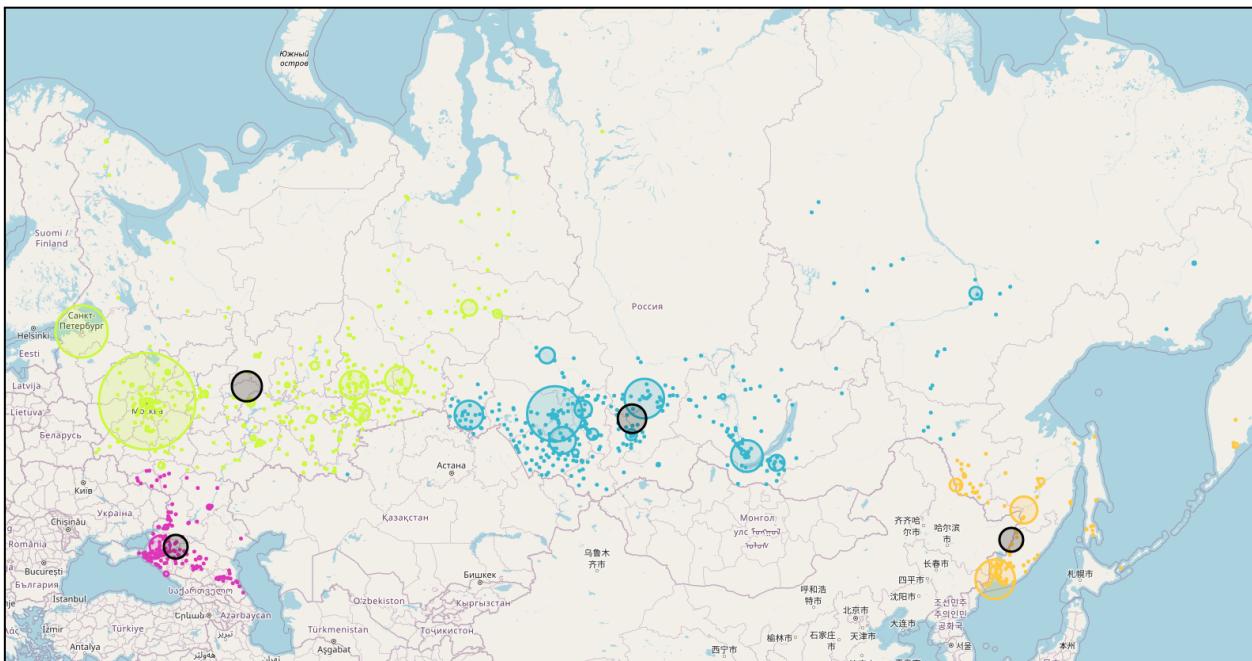


Источник: Составлено автором.

Рисунок 15 – Совместный график Elbow Curve и GAP

для пространственной кластеризации

Визуализация построения кластеризации по четырем центроидам для пространственных стороны имеющихся данных выглядит следующим образом (рис. 16).



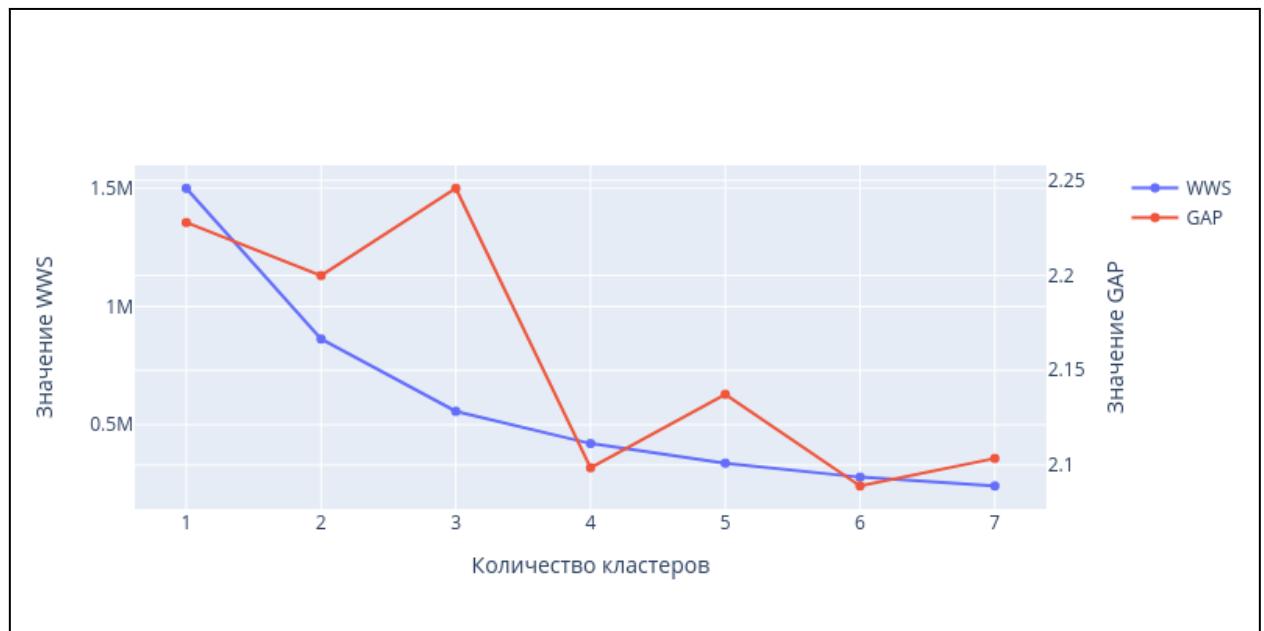
Источник: Составлено автором.

Рисунок 16 – Круговая карта пространственной кластеризации

центров продаж автомобилей с сайта Drom.ru

Как видно, четко выделяются следующие области – дальний восток, Сибирь, северо-западная часть России и Северный Кавказ, которые можно назвать несколько атомарными между собой.

Валидация результатов кластеризации для определения ценовых сегментов, при использовании логарифма цены и объема двигателя, дает следующие кривые WWS и GAP (рис. 17).



Источник: Составлено автором.

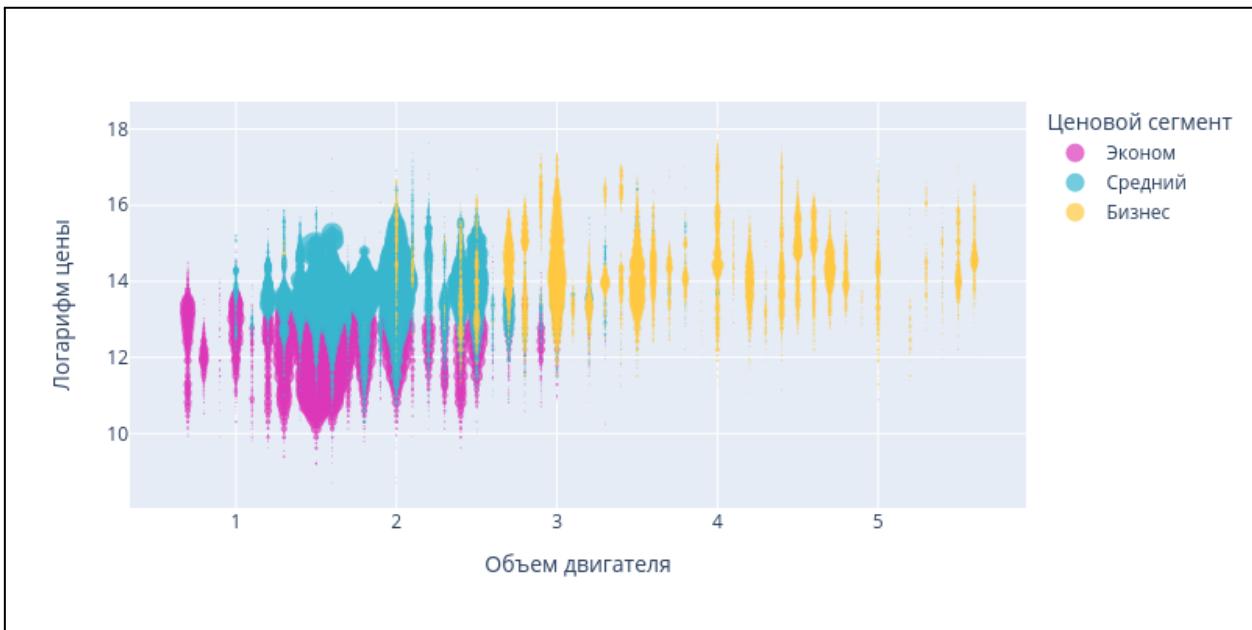
Рисунок 17 – Совместный график Elbow Curve и GAP

для кластеризации по цене и объему двигателя

Как видно, здесь не возникает сомнений в определении оптимального для выборки количества кластеров, данное число равняется 3.

Соотнеся каждое наблюдение к одному из трех, как было описано выше во 2 главе, каждой модели присваивается тот ценовой сегмент, который является модальным для наблюдений с данной моделью.

Результатом данных действий может служить следующий график (рис. 18), который является группирующей по комбинации цены и двигателя диаграммой рассеяния, показывающей количество наблюдений, подпадающих под определенную комбинацию, изменяя размер данной точки.



*Источник: Составлено автором.*

Рисунок 18 – График результатов кластеризации по цене и объему двигателя автомобилей с сайта Drom.ru сгруппированных по модельной принадлежности

### 3.2. Оценка результатов пространственного анализа

Так как наблюдения в различных регионах не могут быть сравнены напрямую в силу в силу своей структуры, необходимо определить однородные межрегиональные величины, которые можно сравнивать между собой. По причине того, что региональное наличие определенных моделей какого-либо бренда может различным, необходимо сгруппировать по ценовым категориям внутри региона средние выставленные в объявлении цены и на основе этих данных рассчитать статистику Морана.

Результатом данных действий является величина статистики Морана I со значением в 0.561 и значимостью менее, чем одна десятитысячная. Что говорит о том, что имеющиеся наблюдения вероятнее всего обладают положительной пространственной корреляции, а значит, регионы по значениям зависимой переменной – логарифма цены, окружены подобными себе.

Представляются оцененные параметры моделей SLM с фиксированными эффектами для разных ценовых сегментов (рис. А1, А2, А3), а также отдельные значения оценок для параметров при вещественных переменных моделей (таб. 1).

Таблица 1 – Оценки коэффициентов при вещественных переменных  
SLM модели с фиксированными эффектами для разных ценовых сегментов

Переменная	Бизнес	Средний	Эконом
state_W	4.235774	5.267731	4.565412
power	0.004582	0.005854	0.005017
engine_liters	-0.349894	-0.269463	0.037840
mileage	-0.000001	-0.000003	-0.000002
mileage_sqrt	-0.002011	-0.001178	0.000127
state_population	0.000005	0.000002	0.000015
state_urban_rate	0.000762	-0.001032	-0.003778
state_paved_roads_per_capita	-0.002363	-0.006437	-0.009676
state_buses_per_capita	-0.000049	-0.000224	-0.000239
state_log_average_income	-0.004440	-0.037119	-0.119992
state_log_average_invest_per_capita_5years	-0.010491	0.020481	0.036041

*Источник:* Составлено автором.

Из приведенной таблицы можно заметить, среди большинства параметров при переменных оценки имеют различающиеся значения и знаки, при этом, практически все приведенные переменные являются значимыми на уровне 5% (таб. А1, А2, А3), что говорит о том, что оценивание модели отдельно для каждого ценового сегмента действительно играет значимую роль для объяснения логарифма цены автомобиля, в связи с различающейся структурой ценообразования. Как можно заметить, что наибольшее различие между приведенными значениями оценок существует между бизнес сегментом и эконом сегментом, что соответствует эмпирическим ожиданиям.

### 3.3. Эмпирическое оценивание эффекта гарантий, интерпретация

Отсечение выборки по указанным в предыдущей главе 10 тыс. км. сократило выборку до 83611 наблюдений. Данная выборка имеет следующее распределение наблюдений по ценовым сегментам: 10.06% приходится на эконом сегмент, 79.13% – средний, оставшиеся 10.81% – бизнес сегмент. Можно заметить, что текущее распределение сильно отличается от долей ценовых сегментов первоначальной выборки с пробегом до 400 тыс. км., что является следствием того, что наибольшая доля предложения новых автомобилей автосалонами и специализированными компаниями приходится именно на сегмент автомобилей среднего класса. Общее количество уникальных моделей автомобилей составляет 101, 305 и 291 соответственно.

В результате применения точного мэтчинга по модели автомобиля получаются вышеобозначенные блоки, внутри которых применяется PSM. Блоки были отсортированы по признаку достаточности данных, которым выступило наличие количества наблюдений в каждой группе не менее 30 объявлений. Порог в 30 наблюдений для каждой группы, следовательно, суммарно 60 наблюдений, был выбран эмпирическим путем, учитывая количество используемых в дальнейшем признаков в PSM, а также ограниченность имеющейся выборки.

Так, например, первые пять моделей по общему количеству наблюдений имеют следующие соотношения количества наблюдений для группы воздействия и группы контроля (таб 2.)

Таблица 2 – Первые пять моделей с наибольшим количеством общих наблюдений

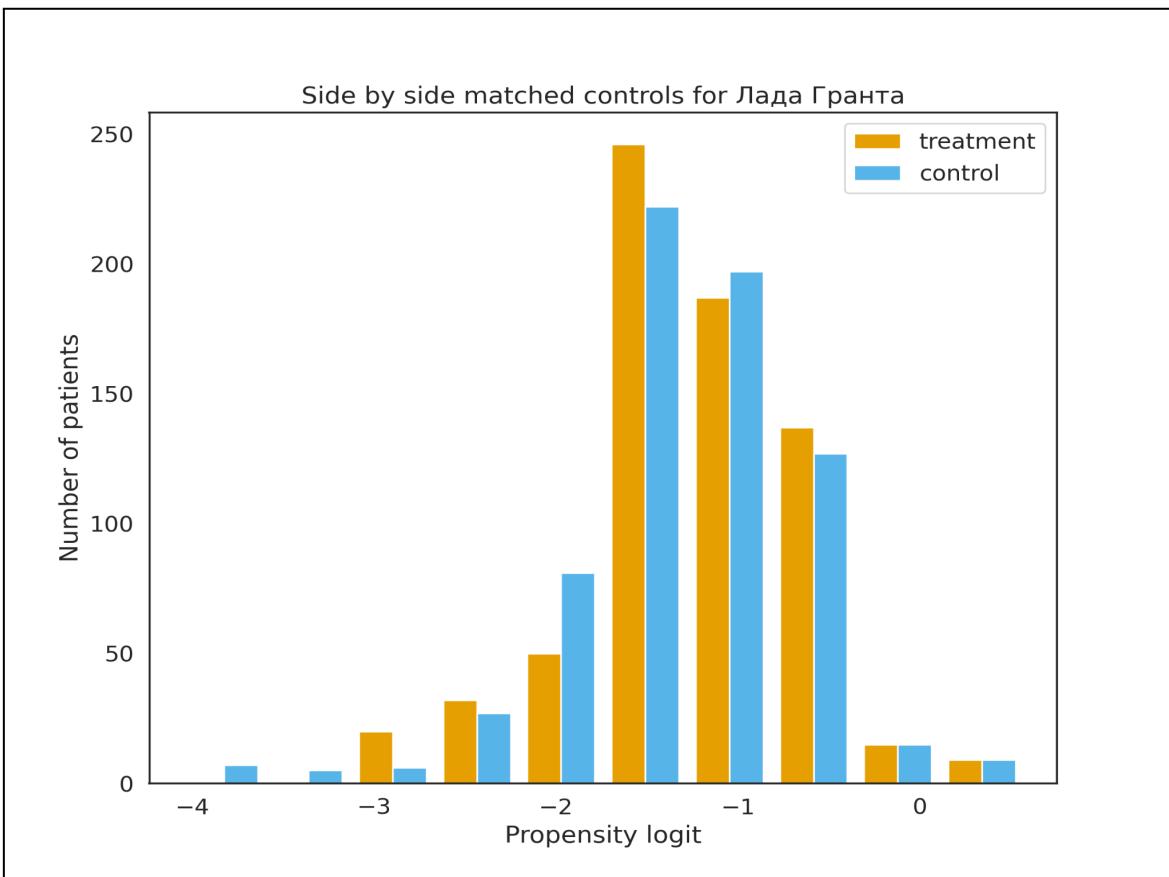
Марка и модель	Ценовой сегмент	Контрольная группа	Группа воздействия	Общее число наблюдений
Лада Гранта	Средний	696	5227	5923
Hyundai Tucson	Средний	901	1636	2537
Kia Sportage	Средний	546	1584	2130
Chery Tiggo 7 Pro	Средний	161	1802	1963
Hyundai Santa Fe	Средний	424	1305	1729

*Источник: Составлено автором.*

Визуализация использования алгоритма Propensity Score Matching в настоящей работе может быть представлена следующим графиком (рис. 19).

Как можно заметить, практически для каждой страты находится соответствие между контрольной группой и группой воздействия, по причине наличия большого количества данных для модели “Лада Гранта” – первой по количеству в выборке.

Избавляясь от воздействия износа вторичных автомобилей, рассчитываются SATE, на основе которых рассчитывается ATE для ценовых сегментов. Так как вышеперечисленные действия производились в логарифмах цен, а соответственно было найдено само отношение цен, необходимо потенцировать полученные эффекты, для того, чтобы получить эффект в процентах.



*Источник:* Составлено автором.

Рисунок 19 – График результатов использования Propensity Score Matching  
для автомобилей бренда “Лада” модели “Гранта”

Результирующая таблица (таб. 3), которая содержит в себе основную информацию о скользуированных эффектах гарантii для различных ценовых сегментов выглядит следующим образом.

Таблица 3 – Рассчитанный эффект гарантii для ценовых сегментов

Ценовой сегмент	Средняя цена, руб.	Средний объем двигателя, л	Количество изначальных наблюдений	Количество наблюдений из мэтчинга	Эффект гарантii (ATE)	Станд. откл.
Эконом	279983.99	1.64	249444	390	<b>87.36%</b>	47,44%
Средний	1354045.75	1.87	416673	13237	<b>107.24%</b>	17.82%
Бизнес	2182329.43	3.64	84061	2863	<b>114.07%</b>	24.13%

*Источник:* Составлено автором.

Полученный эффект гарантii имеет достаточно простую интерпретацию – это величина в которую увеличивается стоимость автомобиля при наличии гарантii со стороны продавца.

Как видно, прослеживается очевидная взаимосвязь между классом автомобиля и его эффектом гарантiiй. ATE для эконом сегмента является отрицательным, что можно снести на факт малого количества наблюдений, полученных в результате мэтчинга и их высокого стандартного отклонения для ATE, поэтому, данный показатель может быть нерепрезентативным. Малое количество наблюдений объясняется тем, что для эконом сегмента присутствует очень малое количество автомобилей без пробега, так как минимальная стоимость нового автомобиля от официального дилера приходится на Lada Granta [Сайт официального дилера..., 2023] со стоимостью более 650 тыс. рублей.

Вместе с этим, стоит учитывать и тот факт, что воздействие износа автомобиля в процессе увеличения пробега достаточно хорошо улавливалось, так как был взят не только сам пробег, но и его корень, который хорошо описывает кривую изменения цены при изменение пробега (рис. 14) на промежутке до 400 тыс. километров. Поэтому, определенная доля эффекта гарантiiй, которая закладывается в первые километры пробега, могла учитываться в величине, которая была прибавлена к стоимости вторичного автомобиля. Из-за чего при нахождении отношения цен сопоставляемых автомобилей цена автомобиля без гарантiiй могла перевесить стоимость автомобиля с гарантiiями.

Величина эффекта гарантiiй для среднего ценового сегмента составляет приблизительно 107% и при этом имеет достаточно низкое стандартное отклонение, что можно снести на высокое количество наблюдений, полученных в результате применения мэтчинга, которое может в силу асимптотичности давать достаточно близкое к генеральной выборке эмпирическое распределение.

Наибольшее значения эффекта гарантiiй принадлежит бизнес сегменту, что является ожидаемым результатом и достаточно точно отражает человеческие представления.

## **ЗАКЛЮЧЕНИЕ**

Проведя описательный анализ автомобильного рынка России и его динамичного развития, удалось выделить актуальную экономически обусловленную проблему, которая может быть решена эффективной законодательной политикой.

Были проанализированы источники первичных микроданных, а также сконструирована программа, позволяющая с асинхронной параллельностью собирать информацию об объявлениях автомобилей на одной из самых крупных площадок России. Собранные данные прошли первичную обработку, а также были проанализированы и очищены.

Сформирована пошаговая схема действий, необходимая для эффективной оценки эффекта гарантiiй для разных ценовых категорий. Определены математические модели и методы, которые нашли теоретическую трактовку.

В результате кластеризации были выявлены 3 основных ценовых сегмента, на которые были распределены имеющиеся модели автомобилей. Построенная эконометрическая модель, учитывающая характеристики автомобиля, региональные эффекты, а также воздействие на цену репутации самой модели. Найдя соответствие наблюдений из контрольной группы и группы воздействия, а также учитывая воздействие износа автомобиля, удалось оценить эффект гарантiiй для разных ценовых категорий автомобилей.

Для ценовых сегментов с большим количеством наблюдений, которые использовались для мэтчинга, итоговое оцененное значение эффекта гарантiiй действительно меньше, чем относительное налоговое бремя, которое несет официально работающая организация, что подтвердило первоначальную гипотезу, которая формировалась в первом пункте первой главы.

Вместе с этим, первоначальное сегментирование имеющейся выборки с помощью кластеризации помогло подтвердить тот факт, что для автомобилей разного класса данный эффект действительно сильно варьируется и играет более значимую роль для более дорогих моделей, что являлось второй гипотезой, которая была сформирована во втором пункте первой главы.

Следствием оценки сегментированного эффекта гарантiiй данного рынка не могут служить прямые очевидные рекомендации для ведения более эффективной налоговой политики, потому что необходимо учитывать разнонаправленные макроэкономические эффекты данных действий, как, например, возможный рост рыночной концентрации на рынке вторичных автомобилей и следующий из этого рост цен на них, который может привести к росту

производства новых автомобилей, изменение налогооблагаемой базы и налоговых поступлений, и другие возможные эффекты.

Однако, результаты данной квалификационной работы могут служить базой и ориентиром для широкого исследования автомобильного рынка Российской Федерации, с помощью которого можно будет составить рекомендации для осуществления оптимальной с точки зрения общества налоговой политики в области автомобильного рынка.

## СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

Авто Авто – платформа №1 для покупки и продажи автомобилей // Аналитическое агентство АВТОСТАТ, URL: <https://www.autostat.ru/infographics/52133/> (дата обращения: 19.09.2022)

Веб аналитика сайта Drom.ru // Similarweb, поставщик цифровой информации, URL: <https://www.similarweb.com/ru/website/drom.ru/#demographics> (дата обращения: 07.02.2023)

Доходы, не подлежащие налогообложению (освобождаемые от налогообложения) // Налоговый кодекс Российской Федерации, часть вторая, Статья 217, от 05.08.2000 N 117-ФЗ (ред. от 28.04.2023)

Имущественные налоговые вычеты, пп. 1 п. 2 // Налоговый кодекс Российской Федерации, часть вторая, Статья 220, от 05.08.2000 N 117-ФЗ (ред. от 28.04.2023)

Индексы потребительских цен на товары и услуги // Единая межведомственная информационно-статистическая система, URL: <https://www.fedstat.ru/indicator/31074> (дата обращения: 23.02.2023)

Как автобизнес пережил 2022 год: спрос, цены, объём предложений // Журнал Авто.ру, 20.12.2022, URL: <https://auto.ru/mag/article/kak-avtobiznes-perezhil-2022-god/> (дата обращения: 07.01.2023)

Компания Rolls-Royce приостанавливает поставки автомобилей в Россию // Сетевое издание "forbes.ru", 04.03.2022, URL: <https://www.forbes.ru/forbeslife/458081-kompaniya-rolls-royce-priostanavlivaete-postavki-avtomobilej-v-rossiu> (дата обращения: 14.02.2023)

Официальный сайт ООО «Автомобильная статистика», URL: <https://www.autostat.ru/> (дата обращения: 10.10.2022)

Постановление Правительства РФ от 21.12.2019 N 1764 (ред. от 02.12.2022) "О государственной регистрации транспортных средств в регистрационных подразделениях Государственной инспекции безопасности дорожного движения Министерства внутренних дел Российской Федерации"

Приложение к ежегоднику социально-экономических показателей Российской Федерации // Федеральная служба государственной статистики, 30.12.2022, URL: <https://rosstat.gov.ru/folder/210/document/13396> (дата обращения: 14.01.2023)

Решение Коллегии Евразийской экономической комиссии от 22.09.2015 № 122 "Об утверждении Порядка функционирования систем электронных паспортов транспортных средств

(электронных паспортов шасси транспортных средств) и электронных паспортов самоходных машин и других видов техники"

Рынок автомобилей в России изменился до неузнаваемости. Как теперь купить иномарку и кто пришел на смену дилерам? // Российское новостное интернет-издание "Lenta.ru", 27.10.2022, URL: <https://lenta.ru/articles/2022/10/27/avtomobili/> (дата обращения: 16.11.2022)

Сайт официального дилера LADA в Новосибирске "Авто-1", URL: <https://avto1.lada.ru/> (дата обращения: 29.01.2023)

Статистика продаж автомобилей в Российской Федерации // Комитет автопроизводителей Ассоциации европейского бизнеса, URL: <https://abreview.ru/stat/aeb/> (дата обращения: 25.02.2023)

Anderson D. W., Kish L., Cornell R. G., On Stratification, Grouping and Matching // Scandinavian Journal of Statistics, 7 (2), 1980. – 61-66 c.

Arthur D., Vassilvitskii S., K-means++: The Advantages of Careful Seeding // Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007. – 1-9 c.

Berkhin P., Survey of clustering data mining techniques // Technical report, Accrue Software, San Jose, CA, 2002.

BMW Halts Production in Russia and Stops Exports to the Country // The Wall Street Journal, 01.03.2022, URL: [https://www.wsj.com/livecoverage/russia-ukraine-latest-news-2022-03-01/card/bmw-halts-production-i  
n-russia-and-stops-exports-to-the-country-T36AO5SxgtXwEfXJBsZ8](https://www.wsj.com/livecoverage/russia-ukraine-latest-news-2022-03-01/card/bmw-halts-production-in-russia-and-stops-exports-to-the-country-T36AO5SxgtXwEfXJBsZ8) (дата обращения: 14.02.2023)

Chowdhury M., Apon A., Dey K., Data Analytics for Intelligent Transportation Systems // Elsevier, 2017. – 31-67 c.

Dehejia R. H., Wahba S., Propensity Score-Matching Methods for Nonexperimental Causal Studies // The Review of Economics and Statistics, 84 (1), 2002: 151–161 c.

Hausman J.A., Griliches Z., Hall B., Missing Data and Self-selection in Large Panels // Annales de l'INSEE, 30-1, 1978. – 137-176 c.

Heckman J. J., Ichimura H., Todd P. E., Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme // The Review of Economic Studies, 1997. – 605-654 c.

Imposition of Sanctions on 'Luxury Goods' Destined for Russia and Belarus and for Russian and Belarusian Oligarchs and Malign Actors Under the Export Administration Regulations // Department of commerce, Bureau of Industry and Security, 16.03.2022, URL: <https://public-inspection.federalregister.gov/2022-05604.pdf> (дата обращения: 14.02.2023)

Jain A. K. , Dubes R. C., Algorithms for Clustering Data. Prentice-Hall, 1988.

- Kaufman L., Rousseeuw P. J., Partitioning Around Medoids (Program PAM) // Wiley Series in Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons, Inc., 1990. – 68-125 c.
- Ketchen D. J., Shook C. L., The application of cluster analysis in Strategic Management Research: An analysis and critique // Strategic Management Journal, 17 (6), 1996. – 441-458 c.
- King G., Nielsen R., Why Propensity Scores Should Not Be Used for Matching // Political Analysis, 27 (4), 2019. – 435-454 c.
- Kupper L. L., Karon J. M., Kleinbaum D. G., Morgenstern H.; D. K. Lewis, Matching in Epidemiologic Studies: Validity and Efficiency Considerations // Biometrics, 37 (2), 1981. – 271–291 c.
- Li H., Calder C. A., Cressie N., Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model // Geographical Analysis, 39 (4), 2007. – 357–375 c.
- Lloyd S. P., Least squares quantization in pcm // IEEE Transactions on Information Theory, 28(2), 1982. – 129-136 c.
- Maddala G.S., The Use of Variance Components Models in Pooling Cross Section and Time Series Data // Econometrica, 39, 1971. – 341–58 c.
- Moran P. A. P., Notes on Continuous Stochastic Phenomena // Biometrika, 37 (1), 1950. – 17-23 c.
- Rosenbaum P. R., Rubin B. R., The Central Role of the Propensity Score in Observational Studies for Causal Effects // Biometrika, 70 (1), 1983. – 41-55 c.
- Rubin D. B., Matching to Remove Bias in Observational Studies // Biometrics, 29 (1), 1973. – 159-183 c.
- Schneider J., Vlachos M., Fast Parameterless Density-Based Clustering via Random Projections // Association for Computing Machinery, New York: ACM Press, 2013. – 861-866 c.
- Sekhon J., The Neyman–Rubin Model of Causal Inference and Estimation via Matching Methods // The Oxford Handbook of Political Methodology, 2009. – 271-299 c.
- Thorndike R. L., Who Belongs in the Family? // Psychometrika, 18 (4), 1953. – 267–276 c.
- Tibshirani R., Walther G., Hastie T., Estimating the number of clusters in a data set via the gap statistic // Stanford University, 2001. – 441-423 c.
- Ukraine: EU agrees fourth package of restrictive measures against Russia // European Commission, 15.03.2022, URL: [https://ec.europa.eu/commission/presscorner/detail/en/ip\\_22\\_1761](https://ec.europa.eu/commission/presscorner/detail/en/ip_22_1761) (дата обращения: 14.02.2023)

Volkswagen в РФ прекратил отгрузки машин дилерам внутри страны // Сетевое издание "Интерфакс.ру", 28.02.2022, URL: <https://www.interfax.ru/business/825159> (дата обращения: 14.02.2023)

Zhao Y., Zhou X., K-means Clustering Algorithm and Its Improvement Research // Journal of Physics: Conference Series, 2021. – 1-4 с.

## ПРИЛОЖЕНИЕ А

Оценки модели SLM с фиксированными эффектами для различных ценовых категорий

	coef	std err	t	P> t	[0.025	0.975]
state_W	4.5654	0.235	19.391	0.000	4.104	5.027
power	0.0050	4.92e-05	102.040	0.000	0.005	0.005
engine_liters	0.0378	0.007	5.602	0.000	0.025	0.051
mileage	-2.061e-06	4.13e-08	-49.868	0.000	-2.14e-06	-1.98e-06
mileage_sqrt	0.0001	2.86e-05	4.446	0.000	7.1e-05	0.000
state_population	1.495e-05	7.13e-07	20.960	0.000	1.36e-05	1.63e-05
state_urban_rate	-0.0038	0.000	-34.344	0.000	-0.004	-0.004
state_paved_roads_per_capita	-0.0097	0.000	-22.746	0.000	-0.011	-0.009
state_buses_per_capita	-0.0002	5.67e-05	-4.211	0.000	-0.000	-0.000
state_log_average_income	-0.1200	0.006	-19.071	0.000	-0.132	-0.108
state_log_average_invest_per_capita_5years	0.0360	0.002	16.019	0.000	0.032	0.040
gearbox_автомат	0.0463	0.006	7.605	0.000	0.034	0.058
gearbox_вариатор	0.2375	0.003	71.022	0.000	0.231	0.244
gearbox_механика	-0.0428	0.002	-21.600	0.000	-0.047	-0.039
gearbox_робот	0.1169	0.005	21.853	0.000	0.106	0.127
drive_4WD	0.0448	0.003	17.098	0.000	0.040	0.050
drive_задний	0.0792	0.003	25.013	0.000	0.073	0.085
wheel_правый	-0.1020	0.002	-50.909	0.000	-0.106	-0.098
body_type_джип/suv 3 дв.	-0.0555	0.009	-5.977	0.000	-0.074	-0.037
body_type_джип/suv 5 дв.	0.0234	0.003	6.824	0.000	0.017	0.030
body_type_купе	0.0293	0.010	2.962	0.003	0.010	0.049
body_type_лифтбек	-0.0373	0.005	-7.534	0.000	-0.047	-0.028
body_type_универсал	0.0093	0.004	2.654	0.008	0.002	0.016
body_type_хэтчбек 3 дв.	-0.0779	0.006	-13.245	0.000	-0.089	-0.066
body_type_хэтчбек 5 дв.	0.0043	0.002	2.028	0.043	0.000	0.008
color_бежевый	-0.0379	0.006	-6.177	0.000	-0.050	-0.026
color_бордовый	2.766e-05	0.006	0.004	0.997	-0.012	0.012
color_голубой	-0.0461	0.006	-7.207	0.000	-0.059	-0.034
color_желтый	-0.0787	0.011	-7.064	0.000	-0.101	-0.057
color_зеленый	-0.0320	0.004	-8.253	0.000	-0.040	-0.024
color_золотистый	-0.0395	0.010	-4.061	0.000	-0.059	-0.020
color_коричневый	0.1736	0.007	25.134	0.000	0.160	0.187
color_красный	-0.0460	0.004	-10.995	0.000	-0.054	-0.038
color_оранжевый	-0.0184	0.015	-1.245	0.213	-0.047	0.011
color_розовый	-0.1203	0.020	-5.974	0.000	-0.160	-0.081
color_серебристый	0.0349	0.003	12.288	0.000	0.029	0.040
color_серый	0.0210	0.002	8.490	0.000	0.016	0.026
color_синий	-0.0226	0.003	-6.799	0.000	-0.029	-0.016
color_фиолетовый	-0.0843	0.009	-8.885	0.000	-0.103	-0.066
color_черный	0.0591	0.003	20.857	0.000	0.054	0.065
engine_type_дизель	0.0148	0.005	2.728	0.006	0.004	0.025
engine_gas_equipment	0.0222	0.006	3.956	0.000	0.011	0.033
engine_hybrid	0.0268	0.009	2.947	0.003	0.009	0.045

*Источник:* Составлено автором.

Рисунок А1 - Оценки модели SLM с фиксированными эффектами для эконом сегмента

	coef	std err	t	P> t	[0.025	0.975]
state_W	5.2677	0.132	39.784	0.000	5.008	5.527
power	0.0059	2.84e-05	205.939	0.000	0.006	0.006
engine_liters	-0.2695	0.003	-99.890	0.000	-0.275	-0.264
mileage	-2.523e-06	2.08e-08	-121.341	0.000	-2.56e-06	-2.48e-06
mileage_sqrt	-0.0012	1.18e-05	-100.129	0.000	-0.001	-0.001
state_population	2.115e-06	3.75e-07	5.648	0.000	1.38e-06	2.85e-06
state_urban_rate	-0.0010	6.77e-05	-15.249	0.000	-0.001	-0.001
state_paved_roads_per_capita	-0.0064	0.000	-25.201	0.000	-0.007	-0.006
state_buses_per_capita	-0.0002	3.51e-05	-6.382	0.000	-0.000	-0.000
state_log_average_income	-0.0371	0.004	-10.121	0.000	-0.044	-0.030
state_log_average_invest_per_capita_5years	0.0205	0.001	15.678	0.000	0.018	0.023
gearbox_автомат	0.0195	0.003	6.507	0.000	0.014	0.025
gearbox_вариатор	0.1014	0.001	70.347	0.000	0.099	0.104
gearbox_механика	-0.0829	0.001	-61.590	0.000	-0.086	-0.080
gearbox_робот	0.0379	0.002	20.203	0.000	0.034	0.042
drive_4WD	0.0258	0.001	22.486	0.000	0.024	0.028
drive_задний	-0.0563	0.002	-23.115	0.000	-0.061	-0.052
wheel_правый	-0.1615	0.001	-108.378	0.000	-0.164	-0.159
body_type_джип/suv 3 дв.	-0.0224	0.005	-4.851	0.000	-0.031	-0.013
body_type_джип/suv 5 дв.	-0.0008	0.002	-0.511	0.610	-0.004	0.002
body_type_купе	-0.0083	0.005	-1.792	0.073	-0.017	0.001
body_type_лифтбек	-0.0097	0.002	-5.256	0.000	-0.013	-0.006
body_type_универсал	0.0246	0.002	12.153	0.000	0.021	0.029
body_type_хэтчбек 3 дв.	-0.0158	0.005	-3.103	0.002	-0.026	-0.006
body_type_хэтчбек 5 дв.	0.0045	0.002	2.861	0.004	0.001	0.008
color_бежевый	-0.0533	0.004	-12.642	0.000	-0.062	-0.045
color_бордовый	-0.0758	0.005	-15.418	0.000	-0.085	-0.066
color_голубой	-0.0694	0.005	-13.705	0.000	-0.079	-0.059
color_желтый	-0.0443	0.010	-4.293	0.000	-0.065	-0.024
color_зеленый	-0.1707	0.004	-42.706	0.000	-0.179	-0.163
color_золотистый	-0.1467	0.007	-20.166	0.000	-0.161	-0.132
color_коричневый	0.0336	0.003	12.531	0.000	0.028	0.039
color_красный	-0.0583	0.003	-22.173	0.000	-0.063	-0.053
color_оранжевый	-0.0020	0.007	-0.306	0.760	-0.015	0.011
color_розовый	-0.0824	0.025	-3.302	0.001	-0.131	-0.033
color_серебристый	-0.0548	0.002	-31.658	0.000	-0.058	-0.051
color_серый	-0.0122	0.001	-8.688	0.000	-0.015	-0.009
color_синий	-0.0372	0.002	-18.820	0.000	-0.041	-0.033
color_фиолетовый	-0.0689	0.006	-11.885	0.000	-0.080	-0.058
color_черный	0.0094	0.001	7.180	0.000	0.007	0.012
engine_type_дизель	0.1075	0.002	54.211	0.000	0.104	0.111
engine_gas_equipment	0.0121	0.004	2.942	0.003	0.004	0.020
engine_hybrid	0.2188	0.002	99.448	0.000	0.214	0.223

Источник: Составлено автором.

Рисунок А2 - Оценки модели SLM с фиксированными эффектами для среднего сегмента

	coef	std err	t	P> t	[0.025	0.975]
state_W	4.2358	0.436	9.717	0.000	3.381	5.090
power	0.0046	3.61e-05	126.908	0.000	0.005	0.005
engine_liters	-0.3499	0.004	-95.148	0.000	-0.357	-0.343
mileage	-1.173e-06	5.6e-08	-20.940	0.000	-1.28e-06	-1.06e-06
mileage_sqrt	-0.0020	3.42e-05	-58.757	0.000	-0.002	-0.002
state_population	5.392e-06	1.11e-06	4.864	0.000	3.22e-06	7.56e-06
state_urban_rate	0.0008	0.000	3.788	0.000	0.000	0.001
state_paved_roads_per_capita	-0.0024	0.001	-3.308	0.001	-0.004	-0.001
state_buses_per_capita	-4.851e-05	0.000	-0.470	0.638	-0.000	0.000
state_log_average_income	-0.0044	0.010	-0.465	0.642	-0.023	0.014
state_log_average_invest_per_capita_5years	-0.0105	0.004	-2.915	0.004	-0.018	-0.003
gearbox_автомат	0.0305	0.007	4.605	0.000	0.018	0.043
gearbox_вариатор	0.0764	0.009	8.746	0.000	0.059	0.093
gearbox_механика	-0.1981	0.008	-25.461	0.000	-0.213	-0.183
gearbox_робот	0.0232	0.013	1.798	0.072	-0.002	0.048
drive_4WD	0.0095	0.004	2.276	0.023	0.001	0.018
drive_задний	-0.0931	0.007	-13.118	0.000	-0.107	-0.079
wheel_правый	-0.1997	0.005	-40.322	0.000	-0.209	-0.190
body_type_джип/suv 3 дв.	-0.1438	0.012	-12.295	0.000	-0.167	-0.121
body_type_джип/suv 5 дв.	0.0180	0.004	4.932	0.000	0.011	0.025
body_type_купе	0.0617	0.012	5.131	0.000	0.038	0.085
body_type_лифтбек	-0.0032	0.008	-0.400	0.689	-0.019	0.012
body_type_универсал	0.0007	0.008	0.090	0.928	-0.015	0.016
body_type_хэтчбек 3 дв.	0.0210	0.016	1.290	0.197	-0.011	0.053
body_type_хэтчбек 5 дв.	-0.0099	0.006	-1.767	0.077	-0.021	0.001
color_бежевый	-0.0878	0.013	-6.918	0.000	-0.113	-0.063
color_бордовый	-0.1296	0.014	-9.296	0.000	-0.157	-0.102
color_голубой	-0.1191	0.023	-5.281	0.000	-0.163	-0.075
color_желтый	-0.1187	0.039	-3.054	0.002	-0.195	-0.043
color_зеленый	-0.2334	0.010	-22.933	0.000	-0.253	-0.213
color_золотистый	-0.1246	0.015	-8.347	0.000	-0.154	-0.095
color_коричневый	-0.0043	0.009	-0.474	0.635	-0.022	0.013
color_красный	-0.1691	0.013	-13.398	0.000	-0.194	-0.144
color_оранжевый	-0.0677	0.030	-2.225	0.026	-0.127	-0.008
color_розовый	-0.4290	0.137	-3.129	0.002	-0.698	-0.160
color_серебристый	-0.1098	0.006	-17.550	0.000	-0.122	-0.098
color_серый	-0.0788	0.005	-15.572	0.000	-0.089	-0.069
color_синий	-0.1401	0.007	-19.787	0.000	-0.154	-0.126
color_фиолетовый	-0.1070	0.032	-3.310	0.001	-0.170	-0.044
color_черный	-0.0476	0.004	-11.961	0.000	-0.055	-0.040
engine_type_дизель	0.2296	0.004	62.638	0.000	0.222	0.237
engine_gas_equipment	0.0099	0.008	1.197	0.231	-0.006	0.026
engine_hybrid	0.0589	0.012	5.002	0.000	0.036	0.082

Источник: Составлено автором.

Рисунок А3 - Оценки модели SLM с фиксированными эффектами для бизнес сегмента