Разведочный анализ данных был проведён с помощью инструмента Apache Spark. Apache Spark – это Big Data фреймворк с открытым исходным кодом для распределённой пакетной и потоковой обработки неструктурированных и слабоструктурированных данных, входящий в экосистему проектов Hadoop.

Spark состоит из следующих компонентов:

1. Apache Spark Core - базовый движок для обработки данных, который лежит в основе всей платформы. Ядро взаимодействует с системами хранения, управляет памятью, планирует и распределяет нагрузку в кластере. Также оно отвечает за поддержку API языков программирования.
2. Spark SQL - модуль служит для упрощения работы со структурированными данными и позволяет выполнять запросы на языке SQL. Главная его задача состоит в том, чтобы дата-инженеры не задумывались о распределенной природе хранения данных, а сосредоточились на сценариях их использования.
3. Streaming - обеспечивает масштабируемую, высокопроизводительную и отказоустойчивую потоковую обработку данных в реальном времени. В качестве источников данных для Spark могут выступать Kafka, Flume, Kinesis и другие системы.
4. MLlib - масштабируемая библиотека машинного обучения низкого уровня. В библиотеке реализованы различные алгоритмы машинного обучения, например кластеризация, регрессия, классификация и коллаборативная фильтрация.
5. GraphX - служит для манипуляций над графами и их параллельной обработки. GraphX может измерять связность графов, распределение степеней, среднюю длину пути и другие показатели. Он также может соединять графы и быстро их преобразовывать. Кроме встроенных операций над графами, в GraphX также есть библиотека с реализацией алгоритма PageRank.

Spark может работать как в среде кластера Hadoop под управлением YARN, так и без компонентов ядра Hadoop, например, на базе системы управления кластером Mesos. Spark поддерживает несколько популярных распределённых систем хранения данных (HDFS, OpenStack Swift, Cassandra, Amazon S3) и языков программирования (Java, Scala, Python, R), предоставляя для них API-интерфейсы.

В качестве исполняемой машины был выбран ноутбук с 8 гигабайтами оперативной памяти и процессором Intel Pentium 4415U с тактовой частотой 2.3 ГГц, и обладающим 2 ядрами и 4 потоками. На машине был запущен docker контейнер и Apache Spark для языка Python в локальном режиме развертывания. В качестве взаимодействия с исходным кодом был выбран инструмент JupyterLab.

В качестве датасета был выбран следующий набор данных - <https://www.kaggle.com/datasets/rupeshraundal/marketcheck-automotive-data-us-canada>. Данный датасет содержит данные о подержанных автомобилях США и Канады за последние 8 лет. Данные собираются с более чем 65 000 веб-сайтов дилеров, чтобы предоставлять наиболее полные и актуальные описания рыночной активности.

Основывая на данных датасета с Kaggle, рассмотрим основные представленные признаки и отберём интересующие нас для дальнейшего анализа.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Признак | Описание | Тип | Комментарий | Включен ли в конечный датасет |
| id | Уникальный идентификатор записи в датасете | Строковый | - | Да |
| vin | Идентификационный номер транспортного средства (англ. Vehicle identification number, VIN) | Строковый | Не требуется, для обозначения уникальности автомобиля существует признак id | Нет |
| price | Цена автомобиля в долларах США | Целочисленный | - | Да |
| miles | Показания одометра автомобиля | Целочисленный | - | Да |
| stock\_no | Инвентарный номер | Строковый | Не требуется, для обозначения уникальности автомобиля существует признак id | Нет |
| year | Год выпуска | Целочисленный | - | Да |
| make | Производитель | Строковый | - | Да |
| model | Модель автомобиля | Строковый | - | Да |
| body\_type | Тип кузова | Строковый | - | Да |
| trim | Версия или конфигурация автомобиля | Строковый | Неинтересен для анализа, для каждого автомобиля может быть свое обозначение конфигурации | Нет |
| vehicle\_type | Тип транспортного средства | Строковый | Не интересен для анализа, неточно разделяет машины на легковые и грузовые (пикапы и фургоны) | Нет |
| drivetrain | Привод автомобиля (полный, передний, задний и т.д.) | Строковый | Неинтересен для анализа | Нет |
| transmission | Тип коробки передач | Строковый | - | Да |
| fuel\_type | Тип потребляемого топлива | Строковый | Неинтересен для анализа | Нет |
| engine\_size | Объём двигателя | Число с плавающей запятой | - | Да |
| engine\_block | Расположение цилиндров в двигателе | Строковый | Неинтересен для анализа | Нет |
| seller\_name | Наименование дилера автомобилей | Строковый | Неинтересен для анализа | Нет |
| street | Улица дилера автомобилей | Строковый | Неинтересен для анализа | Нет |
| city | Город дилера автомобилей | Строковый | Неинтересен для анализа | Нет |
| state | Штат дилера автомобилей | Строковый | Неинтересен для анализа | Нет |
| zip | Почтовый адрес дилера автомобилей | Строковый | Неинтересен для анализа | Нет |

Выбранный датасет в формате CSV был укорочен до размера в 500 мегабайт (ограничения на размер представлены временем исполнения на рабочей машине) и загружен в среду исполнения Apache Spark. После загрузки файл CSV был преобразован в spark dataframe и были убраны неиспользуемые признаки. На рисунке 1 представлена структура преобразованного датасета.

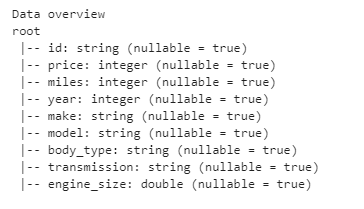


Рисунок 1 – Структура датасета

Общее количество записей в датасете – 2841758.

Рассмотрим датасет на некачественные данные: для этого для каждого признака посчитаем количество записей с отсутствующим или некорректным значением (например, 0 для числовых столбцов). На рисунке 2 представлены результаты подсчета.

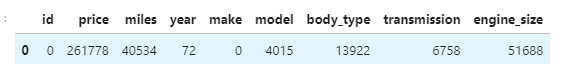


Рисунок 2 – Подсчет некачественных данных

Теперь избавимся от некачественных данных. Около 10 процентов автомобилей в наборе данных не имеют цены, избавимся от таких данных. Так же избавимся от моделей без года выпуска, модели, размера двигателя и трансмиссии, так как эти признаки являются качественными и строго привязаны к действительности (как пример, автомобили марки smart не могут иметь двигатель объёмом более 1.5 литра)

Доля данных без признака цены около 10 процентов в датасете, по этому данные без этого признака тоже были удалены.

Еще раз для каждого признака посчитаем количество записей с отсутствующим или некорректным значением. Как можно заметить на рисунке 3, в датасете больше нет некачественных данных.

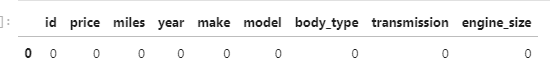


Рисунок 3 – Подсчет некачественных данных после удаления

Рассмотрим статистику числовых признаков датасета (рисунок 4). Можно отчетливо видеть выбросы для каждого из числового признака.

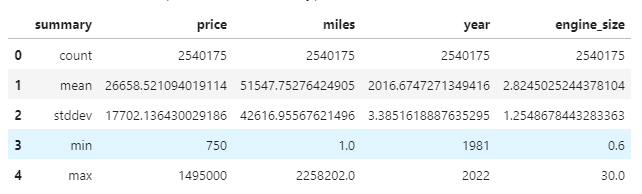


Рисунок 4 – Статистика числовых признаков

Для того что бы избавиться от выбросов данных, воспользуемся межквартильным диапазоном. Межквартильный диапазон, часто сокращенно IQR, представляет собой разницу между 25-м процентилем (Q1) и 75-м процентилем (Q3) в наборе данных. Он измеряет разброс средних 50% значений. Затем на основе межквартильного диапазона рассчитаем пороговое значение для выбросов как 1,5-кратное IQR, вычтем это пороговое значение из 25-го процентиля и добавим его к 75-му процентилю, чтобы получить фактические пределы данных. Данные, не входящие в данный диапазон, удаляются из датасета. Результирующая статистика для всех числовых признаков (за исключением года производства автомобиля) представлена на рисунке 5.

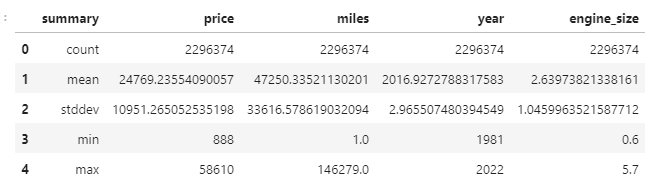


Рисунок 5 – Статистика числовых признаков после удаления выбросов

После этого получившийся датасет сохраняется для дальнейшего использования.

Перейдём к непосредственно к разведочному анализу. Для начала определим самого популярного производителя по объёму автомобилей. Визуализируем 10 из них (рисунок 6)

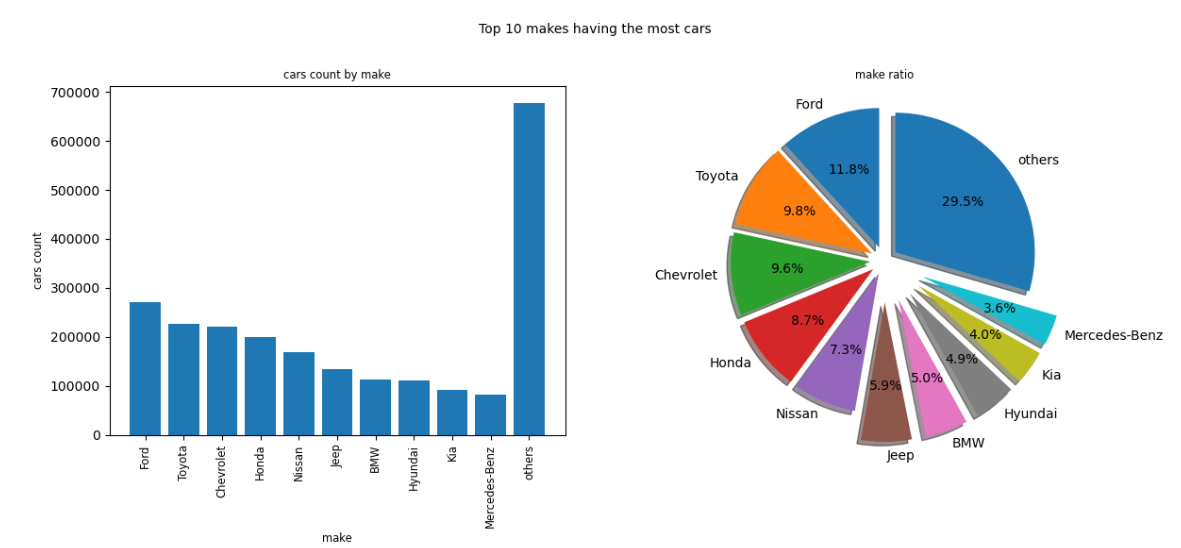


Рисунок 6 – Десять производителей с наибольшим количеством автомобилей (гистограмма и круговая диаграмма)

Можно отметить, что американские производители Ford и Chevrolet вкупе с японским производителем Toyota составляют треть из продаваемых автомобилей.

Рассмотрим распределение десяти производителей с наибольшим количеством автомобилей с точки зрения коробки передач. В датасете признак transmission представлен двумя значениями – Automatic и Manual (автоматическая и ручная коробка передач соответственно). На рисунке 7 представлено процентное соотношение коробок передач для самых популярных производителей.

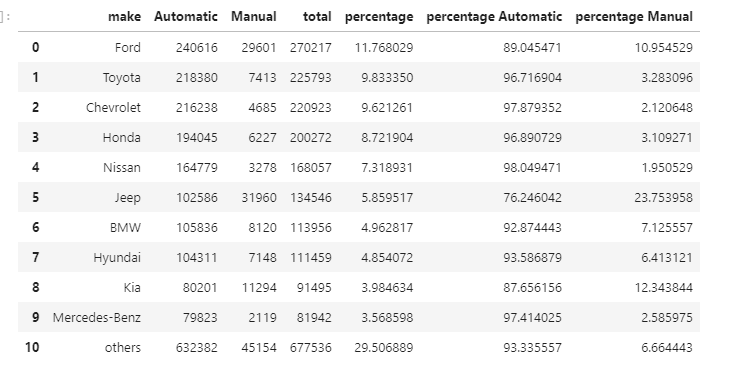


Рисунок 7 – Распределение трансмиссии десяти производителей с наибольшим количеством автомобилей

Среди этих данных можно отметить производителей внедорожников Jeep – почти четверть автомобилей этого производителя оснащены ручной коробкой передач.

Визуализируем производители с наибольшим процентом ручных трансмиссий по отношению к общему числу автомобилей производителя (рисунок 8).

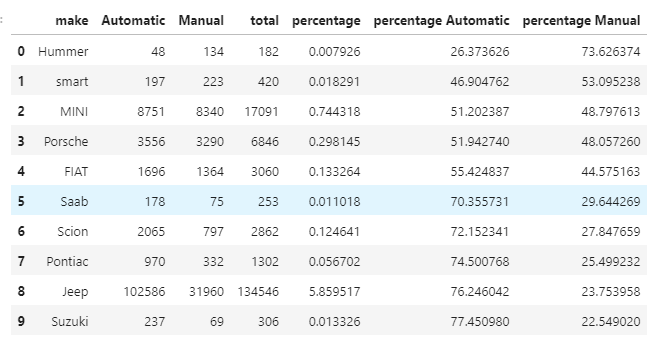


Рисунок 8 – Производители с наибольшим процентом ручных коробок передач у автомобилей

Почти три четверти автомобилей военного производителя внедорожников Hummer оснащены ручной трансмиссией. Так же можно отметить производителя спорткаров Porsche.

Рассмотрим статистику по объему двигателя для десяти самых популярных производителей.

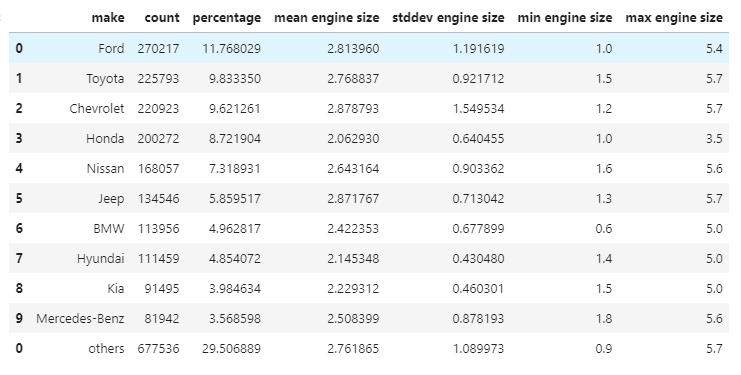


Рисунок 9 – статистика объёма двигателя десяти производителей с наибольшим количеством автомобилей

По данным на рисунке 8 можно сделать вывод, что среднее значение объёмов двигателя находится в диапазоне от 2 до 2.8.

Визуализируем 9 случайных автомобильных производителей с помощью диаграммы размаха (рисунок 10).

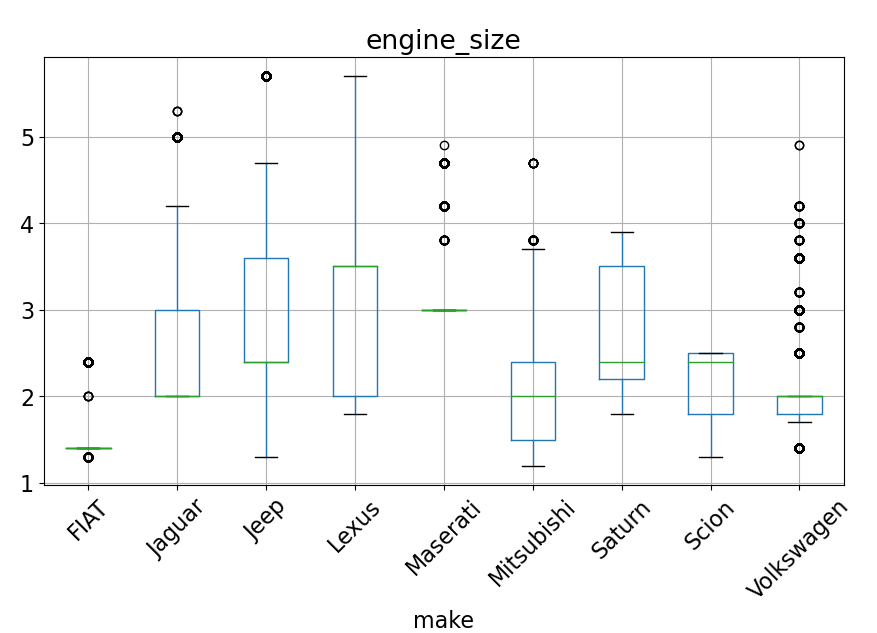


Рисунок 10 – диаграмма размаха объёмов двигателей

Рассмотрим каждую из представленных производителей автомобилей. Производитель Fiat ориентирован на малолитражные автомобили, поэтому средний объём двигателя не превышает 1.4 литра. Модели Jaguar являются спортивными автомобилями с большим объёмом двигателя. Автомобили производителя Jeep являются внедорожниками, которым необходим мощный двигатель. Lexus ориентирован на машины среднего и средне-высокого класса. Maserati является производителем спортивных автомобилей высокого класса, чем и обусловлены выбросы по объёму двигателя. Mitsubishi, Scion производят обыкновенные автомобили. Saturn так же, как и Jeep ориентирован на производство внедорожников. Volkswagen производит большой спектр автомобилей, в том числе фургоны и внедорожники, чем и обусловлены данные выбросы.

Визуализируем распределение цены в виде гистограммы (рисунок 11).

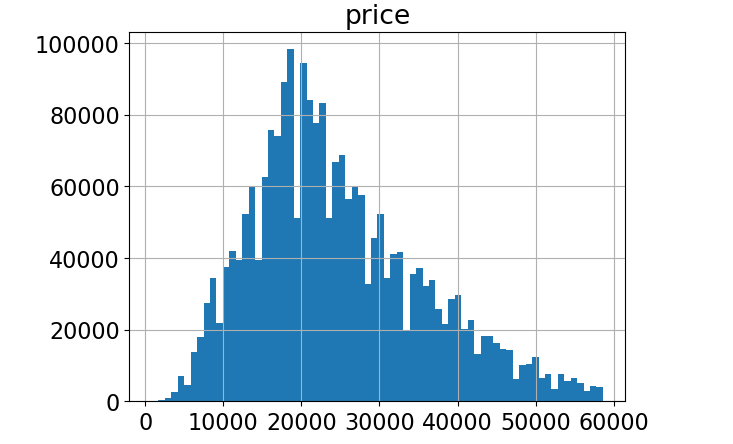


Рисунок 11 – распределение цены автомобилей

Можно отметить, что наибольшей популярностью пользуются автомобили от 17000 до 21000 тысяч долларов.

Рассмотрим, как менялась средняя цена на автомобили за последние пять лет (рисунок 12).

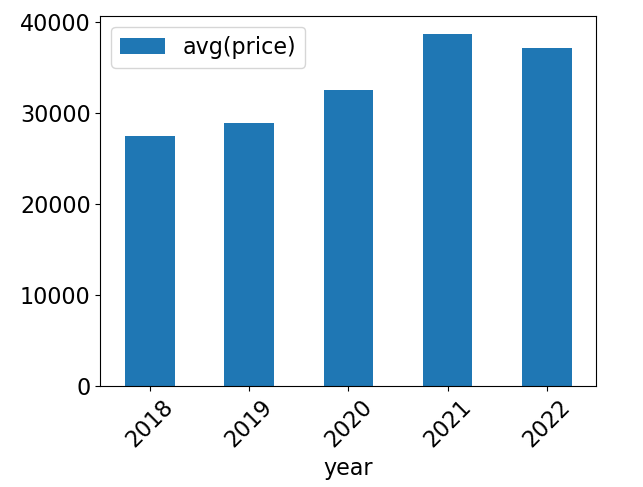


Рисунок 12 – распределение средней цены автомобилей за 2018-2022 года

Можно отметить рост средней цены автомобилей с 2018 по 2021 года.

Рассмотрим самые популярные модели за 2021 и 2022 года (рисунок 13).

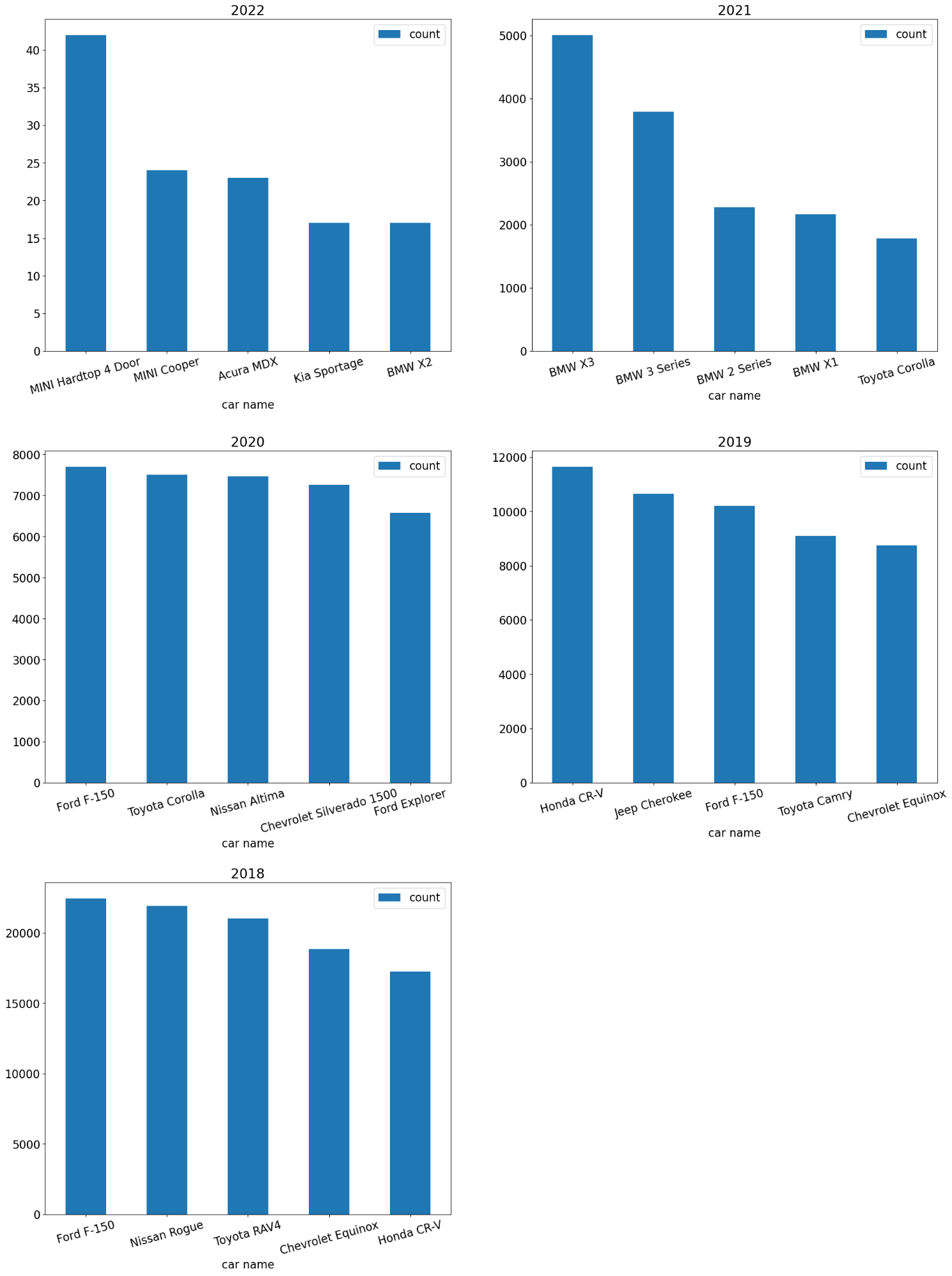


Рисунок 13 – самые популярные автомобили за 2021 и 2022 года

За 2021 года 4 из пяти самых популярных автомобилей принадлежат производителю BMW.

Самым дорогим (по средней цене автомобиля) является немецкий производитель Porsche (рисунок 14).

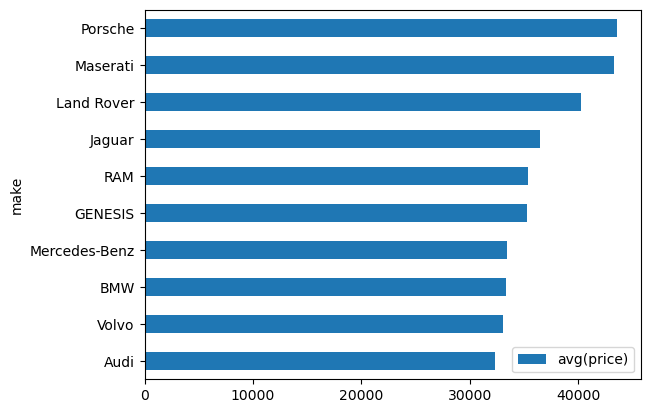


Рисунок 14 – средняя цена автомобилей (первые десять производителей)

Найдем самый популярный автомобиль (рисунок 15).

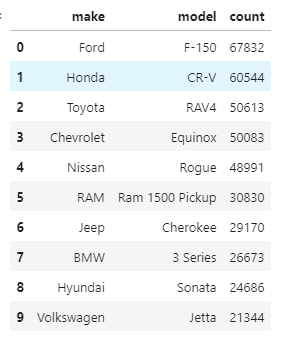


Рисунок 15 – самый популярный автомобиль



Рисунок 16 – Ford F-150 (источник: <https://ru.motor1.com/ford/f150/>)

Визуализируем среднюю цену на эту модель за последнее десять лет (рисунок 17).

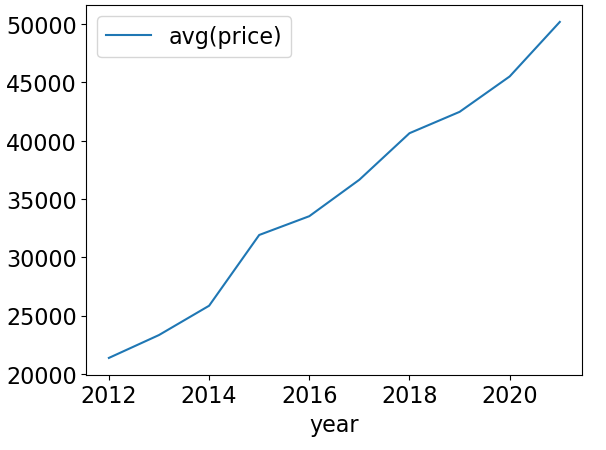


Рисунок 17 –кривая средней цены модели Ford F-150

Рассмотрим новые автомобили Ford F-150. Новыми будем считать те автомобили, пробег которых не превышает 20 тысяч миль. Построим распределение цены для этих автомобилей и нарисуем лучшего распределения (рисунок 18). Линию тренда нарисуем с помощью полиномиальной аппроксимации методом наименьших квадратов с помощью функции numpy.polyfit().

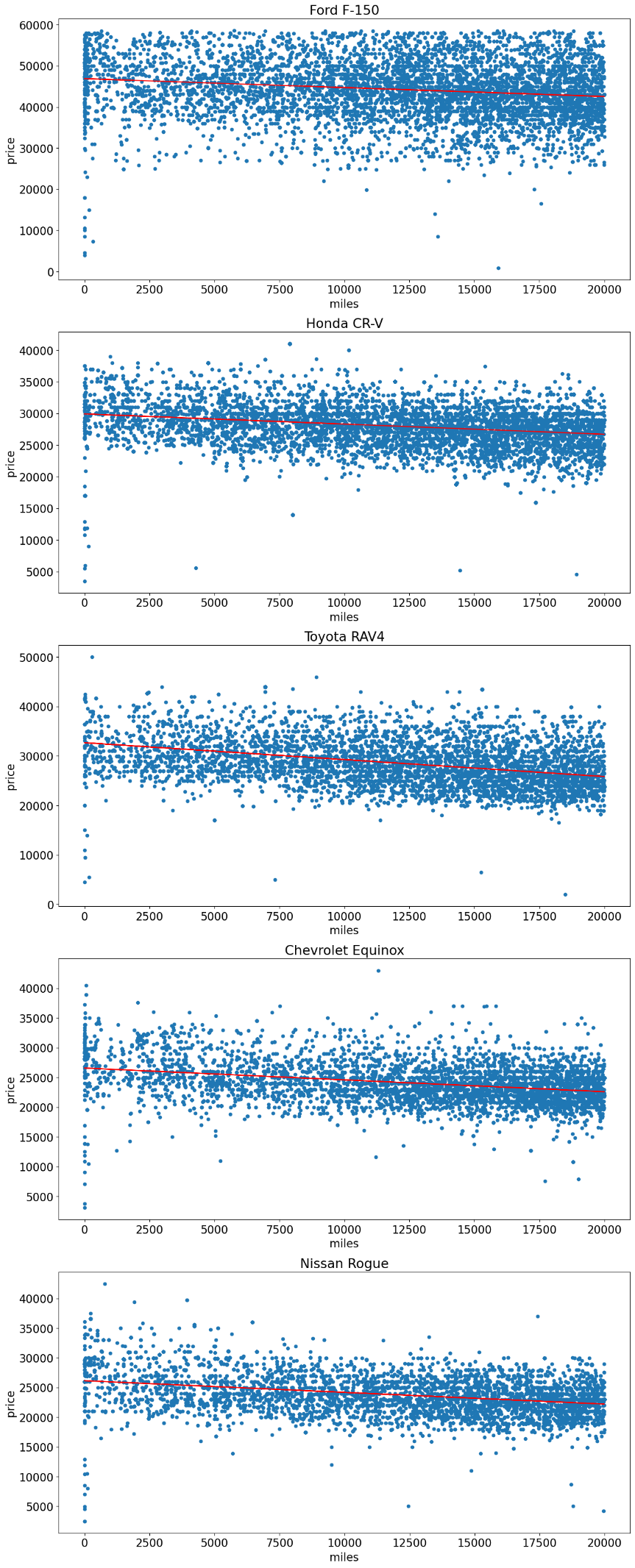


Рисунок 18 – диаграмма рассеяния цены относительно пройдённых миль модели Ford F-150

Из графика можно сделать вывод, сверху и снизу линии лучшего распределения существует разброс в примерно 13 тысяч долларов. Например, вторая по популярности модель Honda CR-V имеет примерно в два раза меньший разброс цены относительно лучшего распределения (рисунок 19).

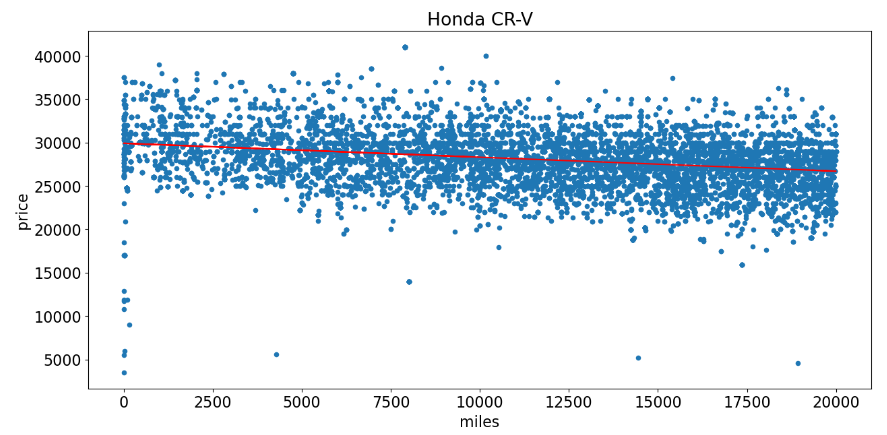


Рисунок 19 – диаграмма рассеяния цены относительно пройдённых миль модели Honda CR-V

Изучим датасет с точки зрения популярности типов кузовов автомобилей.

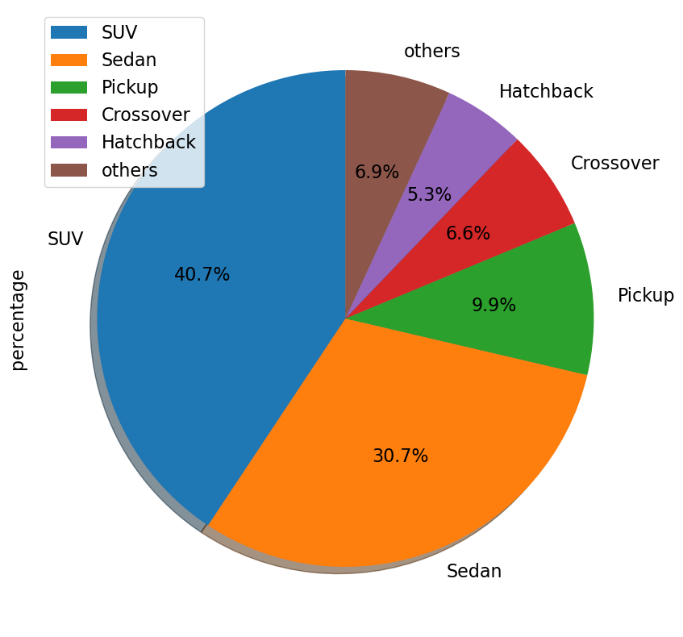


Рисунок 20 – самые популярные типы кузов

Более половины американского рынка составляют полноприводные внедорожные автомобили с типами кузовов SUV и Pickup.

Выводы: был проведен разведочный анализ датасета продаваемых автомобилей в США. Используемый инструмент – Apache Spark. Датасет был исследован на некачественные значения, такие данные были удалены или заполнены.