Часть 2. Данные и их предварительный анализ.

Команда:

Дмитрий Сайфулин, Данат Семенеев, Сергей Петраков, Мария Брусенина.

Тема: Исследование рынка подержанных автомобилей США.

Цель: Оценка количественного влияния и определение значимости географических, технических и визуальных особенностей транспортных средств на ценообразование на рынке б/у автомобилей на территории США.

Гипотезы:

- В штатах с более благоприятной конъюнктурой цены на подержанные авто при прочих равных выше.
- Лучшие состояние авто значимо положительно влияет на цену подержанного авто.
- Пробег значимо отрицательно влияет на цену подержанного авто.

Данные:

В качестве данных мы используем файл vehicles.csv, который был получен нами из интернет ресурса kaggle. На текущий момент это самый большой в мире дата сет подержанных автомобилей, выставленных на продажу на территории США. Важным преимуществом этого дата сета является релевантность данных, а также их количество. Скрапинг новых объявлений происходит раз в несколько месяцев.

Для того, чтобы можно было корректно проводить регрессии мы отчистили данные в Jupyter Notebook от NaN. После всех преобразований наша выборка насчитывает практически 140 тысяч наблюдений.

Комментарий для имплементации кода:

Для корректной работы скрипта в Jupyter Notebook необходимо установить на ваш компьютер дополнительные пакеты. Через Jupyter Notebook это можно сделать командой: *!pip install folium* (если требуется установить folium), также это можно сделать через командную строку командой *pip3 install folium* или *conda install folium*. Скорее всего придётся несколько раз импортировать пакеты. Далее для корректной работы кода рекомендуем файл vehicles.csv расположить в той же папке, где и файл Data_analysis.ipynb.

Summary результатов файла Data_analysis.ipynb.

1. **Описание переменных:**

В изначальном файле vehicles.csv был следующий список переменных со значениями:

ID – входное значение значение ID

url - URL

region craigslist – регион (403 уникальных значения)

price – цена, $

year – год выпуска автомобиля

manufacturer – производитель транспортного средства

Значения: volkswagen, honda, ford, nissan, jeep, gmc, dodge, chrysler, hyundai, chevrolet, toyota, lexus, ram, mini, audi, lincoln, mazda, mercedes-benz, subaru, acura, bmw, cadillac, volvo, buick, saturn, kia, rover, infiniti, mitsubishi, jaguar, mercury, pontiac, alfa-romeo, harley-davidson, tesla, fiat, datsun, land rover, ferrari, aston-martin, porche, morgan, hennessey

model – модель транспортного средства (11268 уникальных значений)

condition – состояние транспортного средства (excellent, good, like new, fair, new, salvage)

cylinders – количество цилиндров, штук.

fuel – тип топлива, подходящего для данного транспортного средства (gas, diesel, other)

odometer – пробег, км.

title – статус транспортного средства (clean, rebuilt, lien, missing, salvage, parts only)

transmission – тип трансмиссии

drive – тип привода (4wd, rwd, fwd)

size – размер транспортного средства (compact, mid-size, full-size, sub-compact)

type generic – тип транспортного средства (hatchback, sedan, truck, coupe, SUV, pickup, wagon, convertible, other, offroad, van, mini-van, bus)

paint_color – цвет (black, grey, white, blue, custom, yellow, silver, red, green, brown, purple, orange)

image_url – URL картинки

description – описание транспортного средства

state – штат, в котором размещена заявка

latitude of listing – координаты заявки

2. Анализ данных

Изначальная статистика отсутствующих значений (NaN) в файле vehicles.csv.

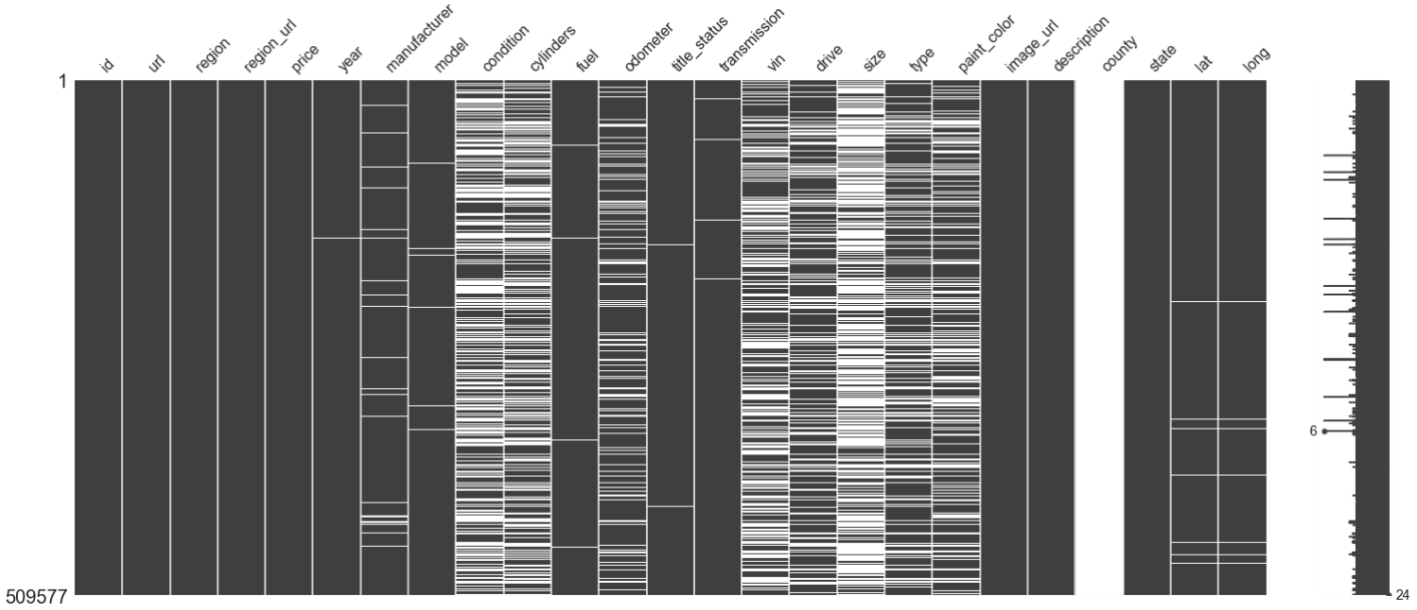

Рис. 1. На данной диаграмме представлена наполненность первоначальных импортируемых данных соответствующими значениями. Белые пропуски соответствуют NaN. От них для корректности мы впоследствии избавились, сократив нашу выборку.

Описательная статистика по данным (в десятках тысяч)

	id	price	year	odometer	lat	long
count	13.996900	13.996900	13.996900	13.996900	13.996900	13.996900
mean	704423.758939	7.292457	0.200915	11.403134	0.003841	-0.009200
std	492.625549	1373.848715	0.000745	12.482957	0.000554	0.001667
min	703256.184100	0.000000	0.191500	0.000000	-0.005596	-0.016594
25%	704083.103500	0.420000	0.200600	6.525800	0.003479	-0.009809
50%	704544.922100	0.840000	0.201000	10.800000	0.003918	-0.008661
75%	704855.082300	1.590000	0.201400	15.050000	0.004238	-0.008019
max	705010.199800	360002.890000	0.202100	1000.000000	0.008157	0.009412

Рис 2. На данной таблице приведена описательная статистика основных значений показателей. Наша выборка содержит 139969 наблюдений. Медианное значение цены составляет 8400$.

Boxplot по ценам.

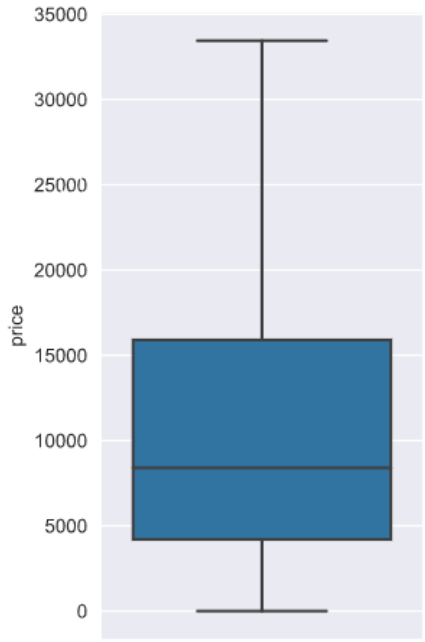

Рис 3. Boxplot - распределение цен (верхняя граница закрашенной фигуры – 3 квартиль, нижняя – 1 квартиль, линия по середине прямоугольника – это медиана).

Диаграмма распределения количества заявок по регионам.

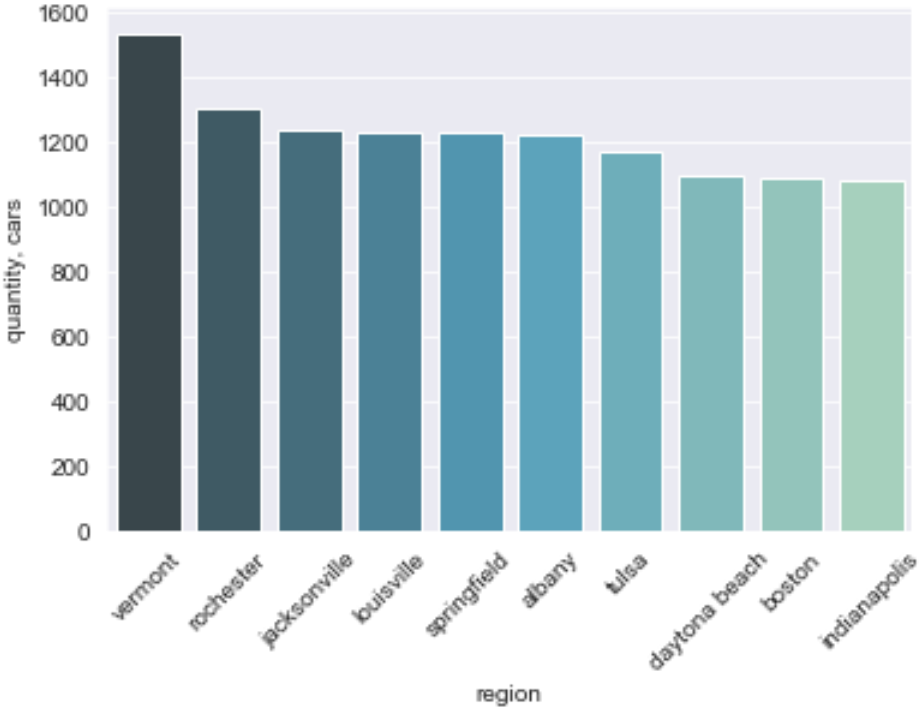

Рис 4. На этой диаграмме демонстрируется распределение предложения по разным регионам США. Слева направо количество заявок уменьшается. Больше всего заявок на продажу подержанных автомобилей люди оставляют в Вермонте, остальные регионы не сильно отличаются между собой по этому показателю.

Диаграмма производства транспортных средств по годам.

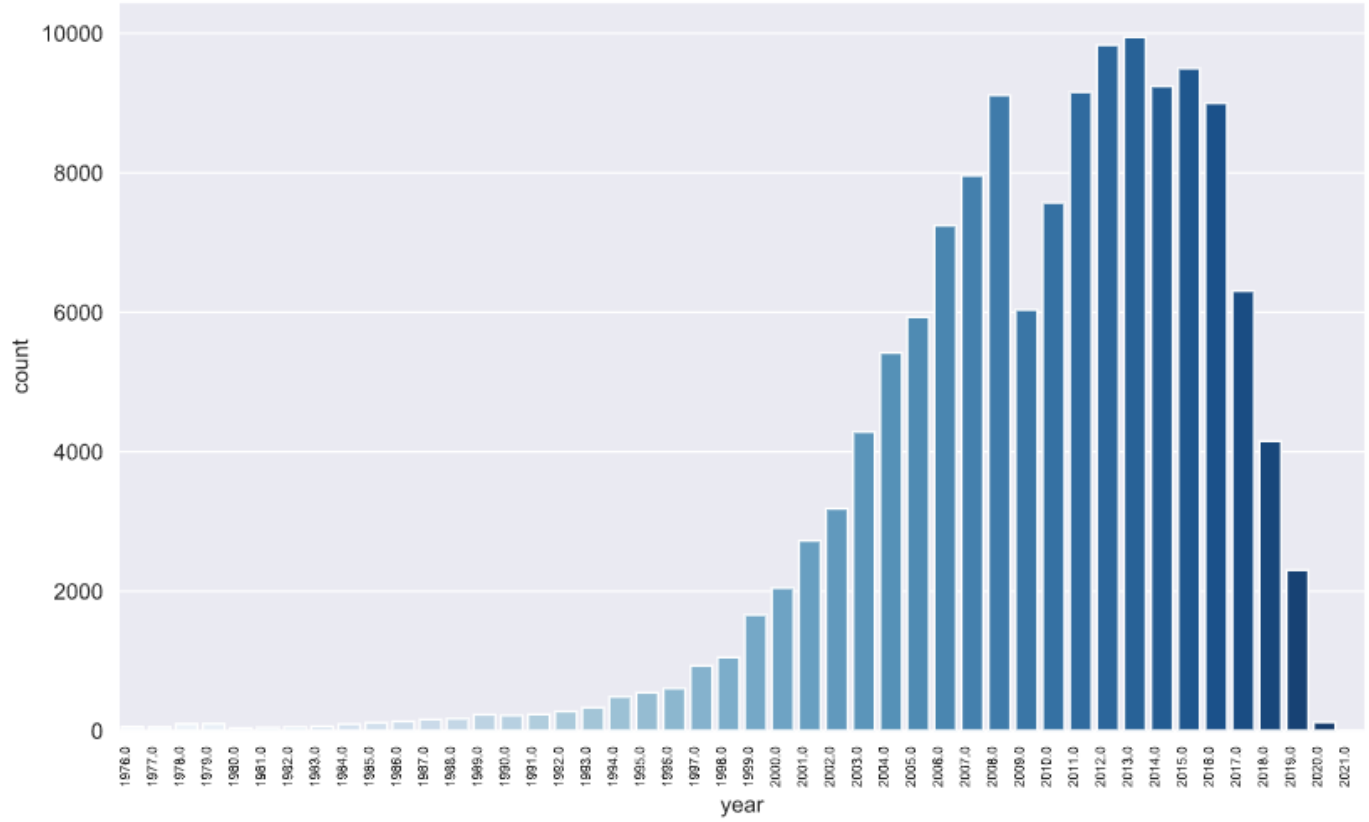

Рис 5. На этой диаграмме показано, сколько машин в какой год было произведено. Особенно заметно влияние Великой рецессии 2008 – 2009 годов. Это отражено в спаде производства в 2009 году.

Количество предложений подержанных машин по производителям.

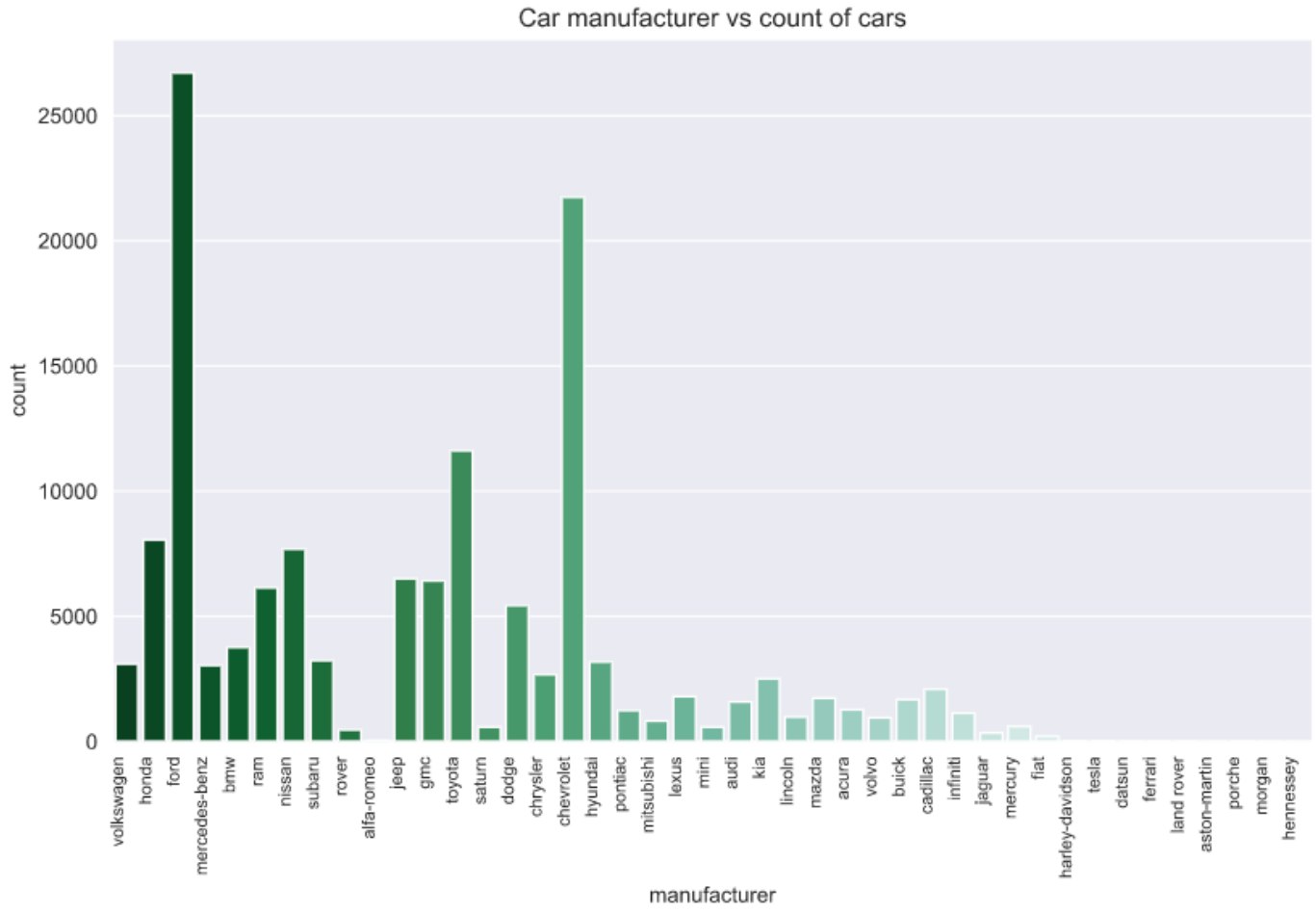

Рис 7. На этой диаграмме иллюстрируются данные о том, какое количество средств транспорта, вставленного на продажу в США какой фирмой произведено. Лидеры: Ford, Chevrolet, Toyota.

Распределение цен по брендам.

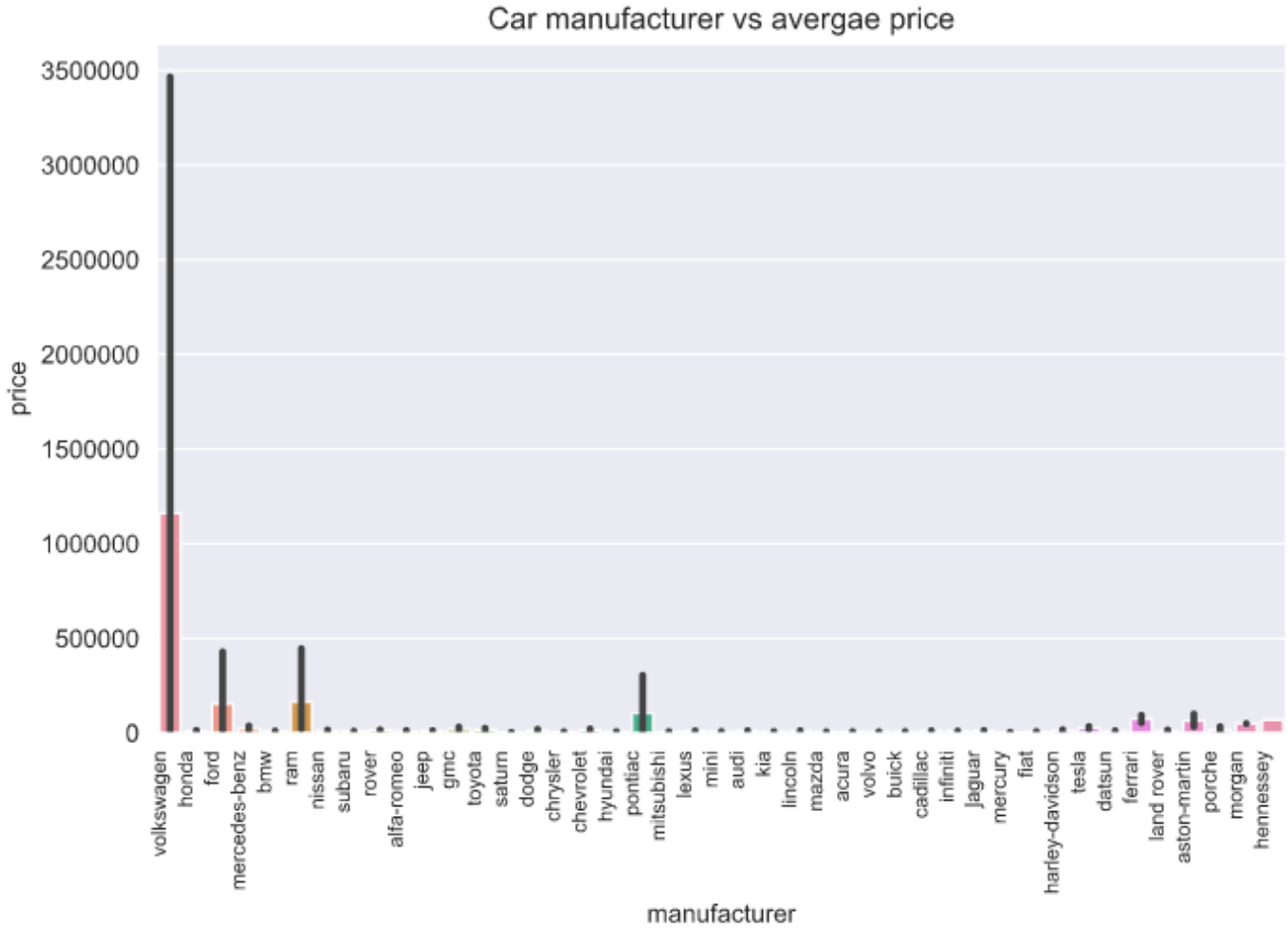

Рис 8. Сильно выделяются крайне высокие цены на Volkswagen, поэтому построим такую е диаграмму, исключив эту марку, а также Ford, Ram и Pontiac, для анализа распределения цен остальных производителей.

Распределение цен на транспортные средства по производителям, за исключением выбросов (Volkswagen, Ford, Ram, и Pontiac).

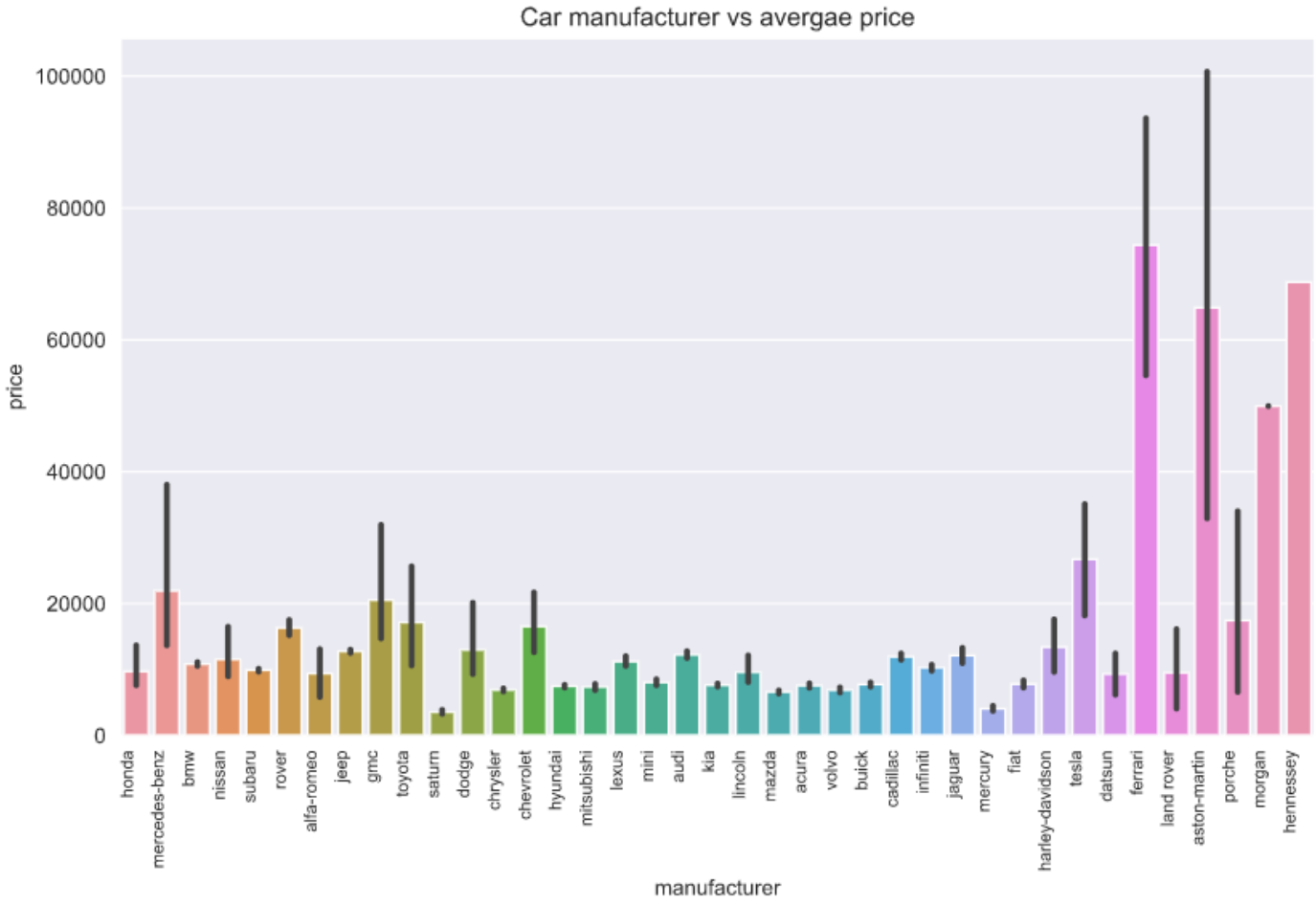

Рис 10. Самые дорогие авто - спорткары (Ferrari, Aston-martin, Porche), технологичные (Tesla) и сравнительно редкие (Morgan, Hennessey). Цены на остальные авто сопоставимы.

Тепловая карта количесвта предлагаемых на рынке машин и их цвета.

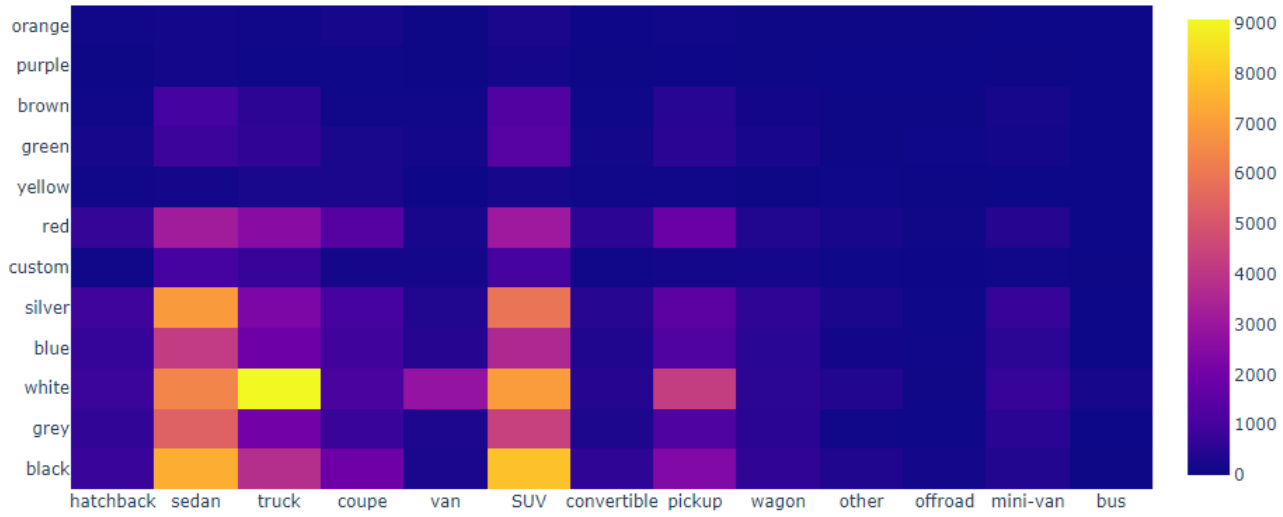

Рис 11. Больше всего машин белого, чёрного и серебристого цвета. Самые распространённые типы: седан, грузовик и внедорожник (SUV).

Тепловая диаграмма, включающая совместную статистику по годам и цветам автомобилей.

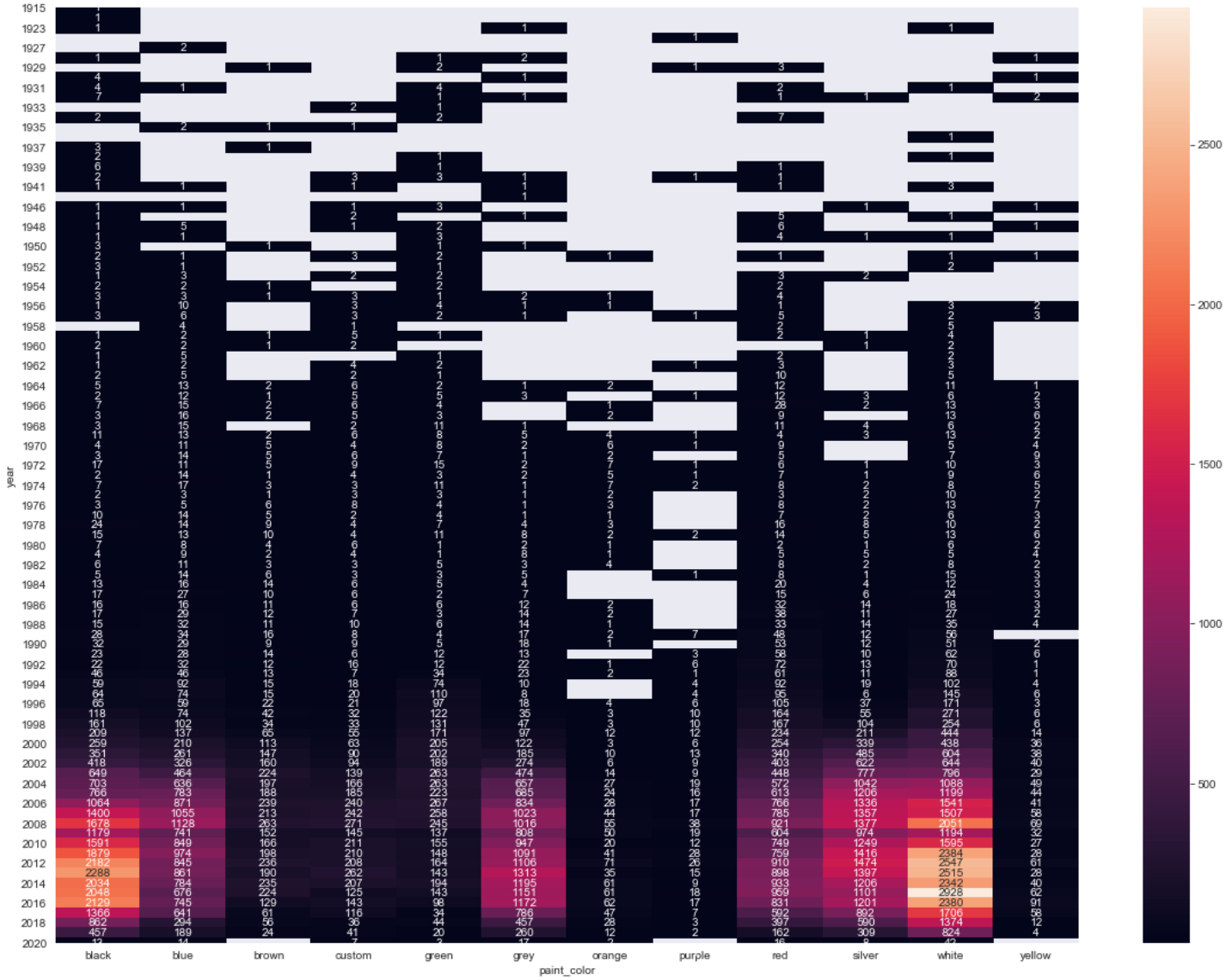

Рис 12. Эта тепловая карта иллюстрирует совместное распределение количества произведённых транспортных средств в разбивке по годам и их цвет. Самые популярные цвета – чёрный и белый, а самые часто встречающиеся года находятся в промежутке между 2004 и 2016.

Статистика состояния автомобилей.

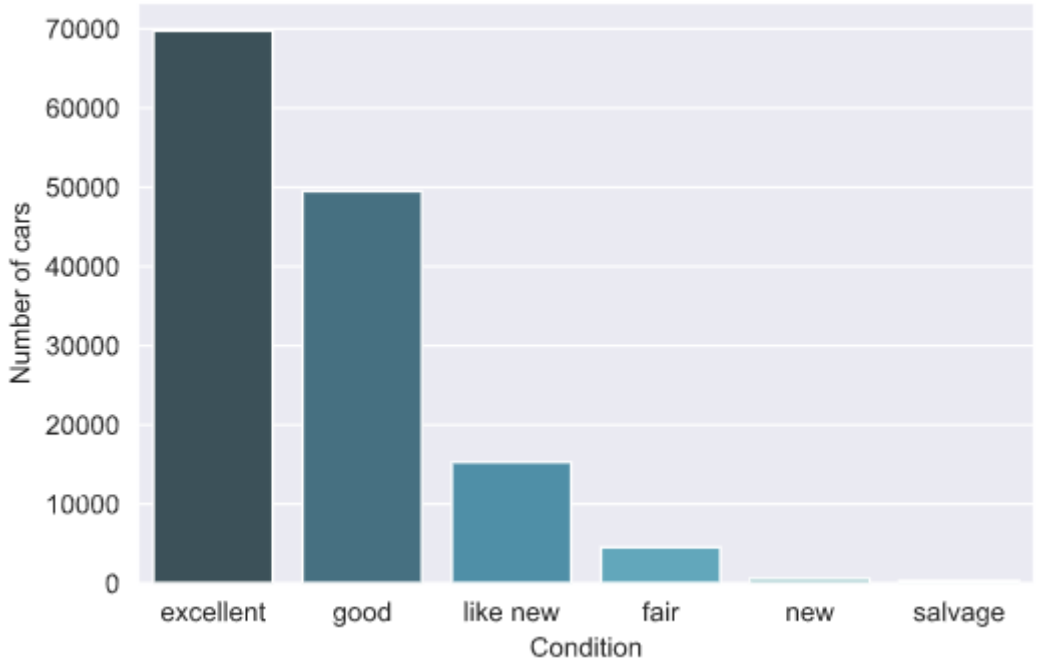

Рис 13. Большая доля автомобилей имеет отличное и хорошее состояние, новые и как новые авто встречаются значительно реже.

Количество автомобилей различных производителей в разбивке по их состоянию.

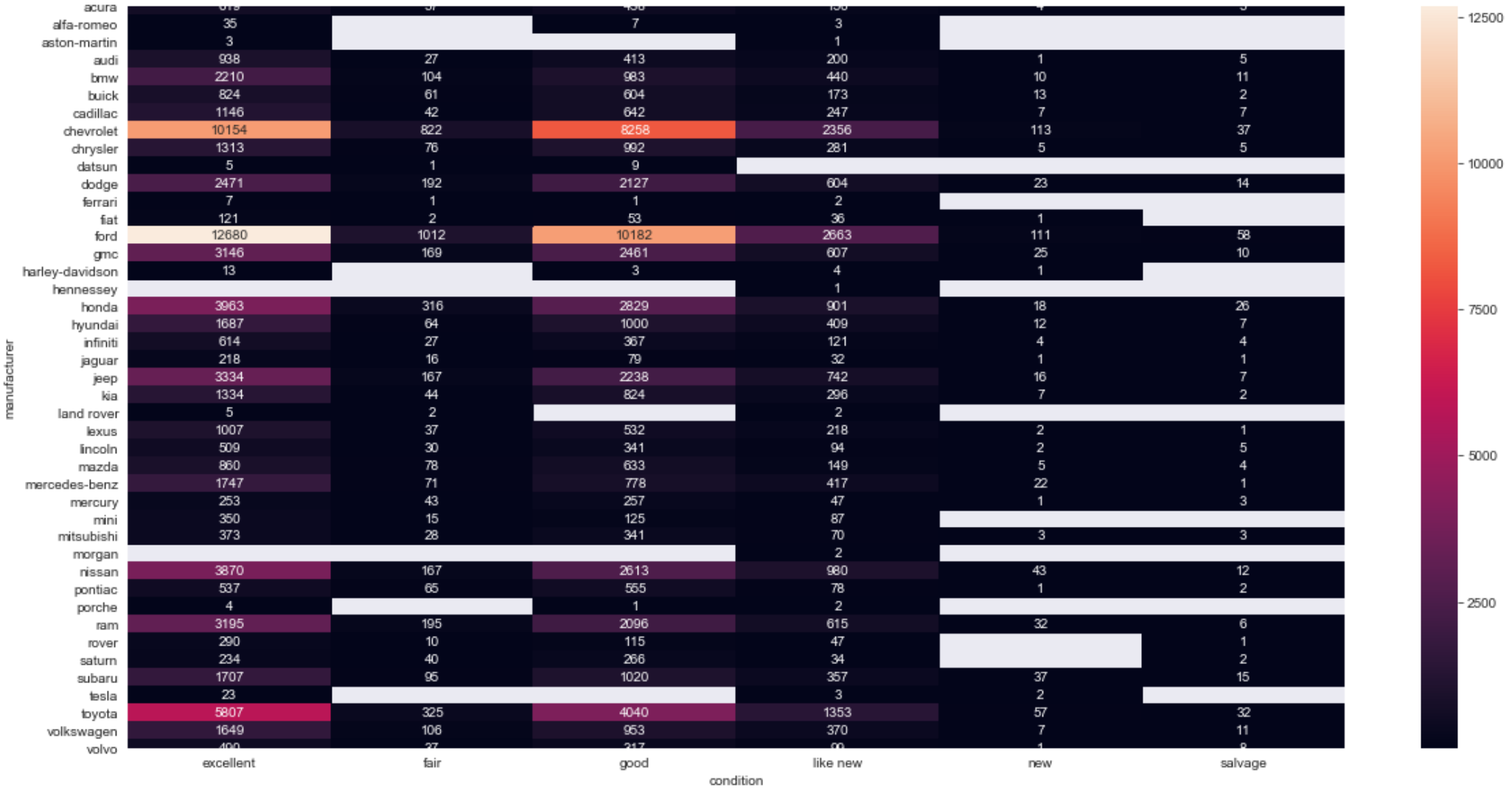

Рис 14. Тепловая карта показывает распределение состояния автомобилей и их производителей. Здесь можно убедиться в том, что в большинстве состояние автомобилей всех марок распределено похожим образом. Белые прямоугольники означают отсутствие данных

Карта предложений подержанных транспортных средств.

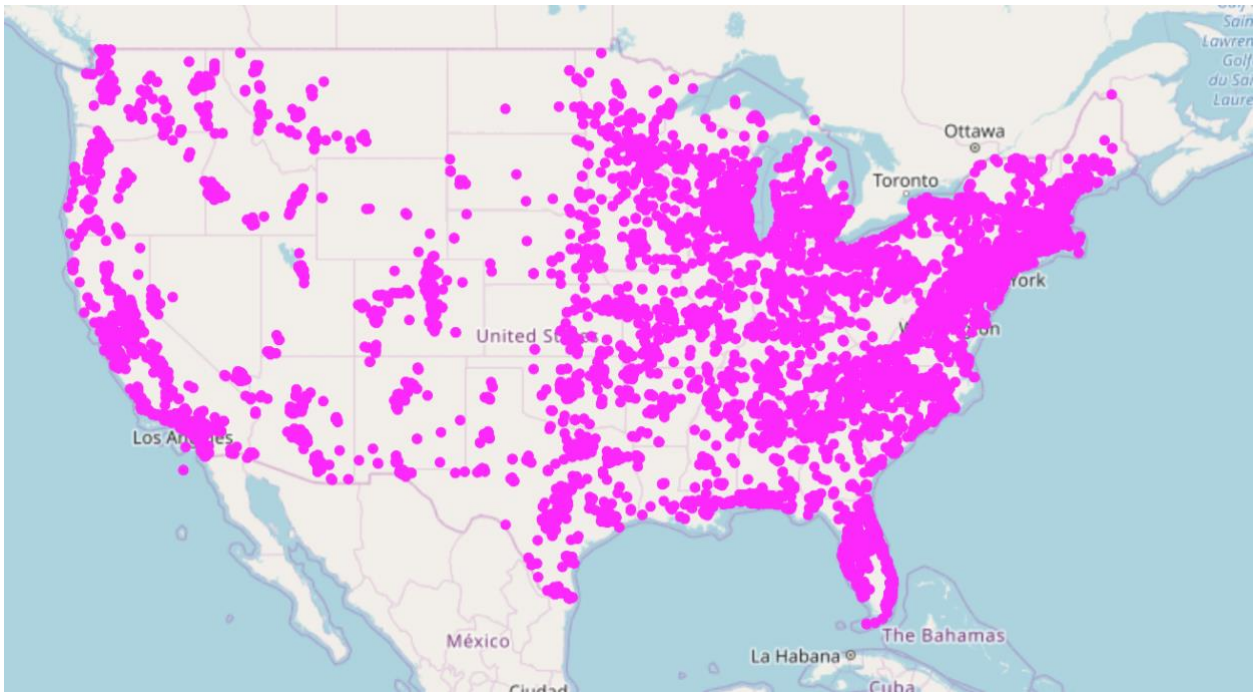

*Рис 15. На этой карте изображено предложение подержанных автомобилей типа **седан**, спроецированных на карту США. Можно отметить увеличение плотности объявлений о продаже в восточной части страны. Однако в целом мы можем говорить о репрезентативности выборки со всей территории страны, так как довольно много наблюдений приходится на центральную и западную часть США.*

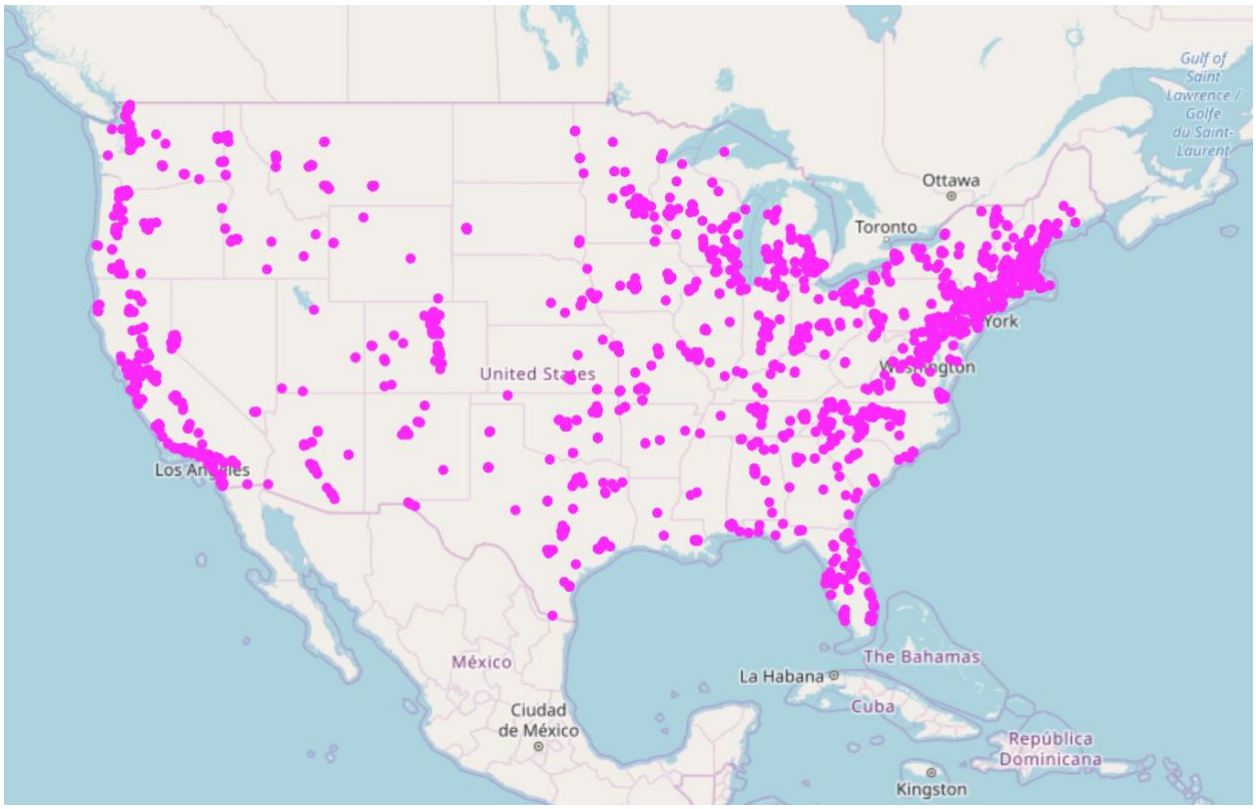

*Рис 16. На этой карте изображено предложение подержанных автомобилей типа **универсал** (wagon), спроецированных на карту США. Также заметны кластеры западной и восточных частей.*

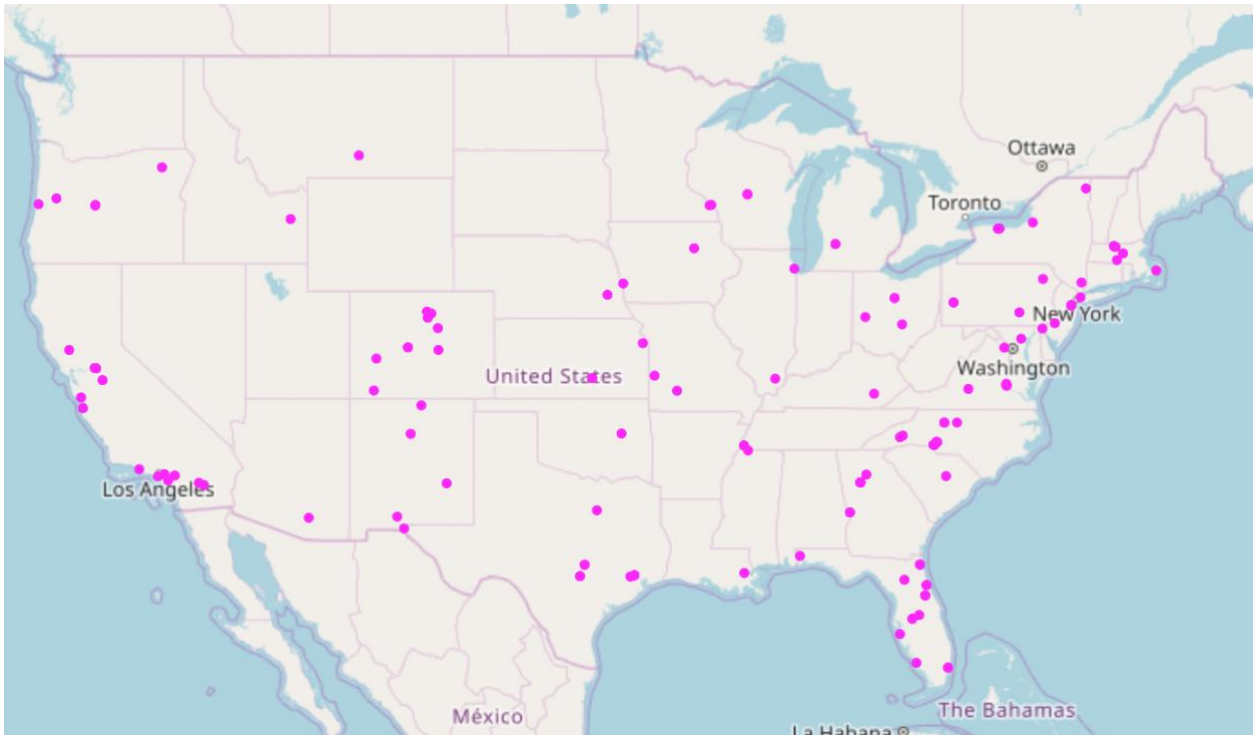

*Рис 17. На этой карте изображено предложение подержанных **автобусов**, спроецированных на карту США. Их количество невелико, однако в восточной части страны автобусов предлагается больше.*