






Исследовательский проект «Ценообразование автомобилей BMW, Audi и Mercedes на площадке drom.ru»

Комаров Максим, Борисенко Григорий

0 датасете



15 620 объявлений 39 моделей Сначала новые объявления

	<p>★ BMW X7, 2023 xDrive40d AT 3.0 л (352 л.с.), дизель, АКПП, 4WD, 50 км без пробега по РФ</p>	<p>17 450 000 ₽ нормальная цена</p>	★
	<p>★ BMW X3, 2019 3.0 л, автомат, 88 110 км</p>	<p>4 449 000 ₽ нормальная цена</p>	★
	<p>★ BMW 7-Series, 2015 730Ld AT xDrive 3.0 л (265 л.с.), дизель, АКПП, 4WD, 154 000 км</p>	<p>3 900 000 ₽ без оценки</p>	★

- файл data.csv
- Собран при помощи веб-скрейпинга
- Содержит информацию более чем о 15.000 объявлениях о продаже
- Включает только автомобили марок BMW, Mercedes и Audi.

О датасете

Контекст: в нашей стране из-за обостренной политической ситуации наблюдаются проблемы с поставками зарубежных авто. В связи с этим цены на автомобили сильно выросли за последние пару лет.

Цель: установить какие характеристики имеют влияние на цену автомобиля. Научиться относить автомобиль к одной из представленных марок по его характеристикам. Разбить авто на кластеры.

- Размер 15029 rows x 22 columns
- Всего 22 переменных; содержат информацию о состоянии автомобиля, ПТС, VIN, технических характеристиках, предыдущих владельцев, модели, комплектации и городе размещения объявления о продаже.
- Источник: drom.ru

Основные переменные

Название	Единицы измерения	Тип	Тип данных	Число пустых значений	Среднее/мода	std
price	рубль	метр.	int64	0	5 445 201	5 926 867
year	год	интерв.	int64	0	2019	-
owner_count	Кол-во владельцев	метр.	float64	4081	3	2
engine_volume_litters	литр	метр.	float64	236	2.5	0.8
power_in_hp	л. с.	метр.	float64	210	245	107
mileage	км	метр.	float64	218	121 013	107 120

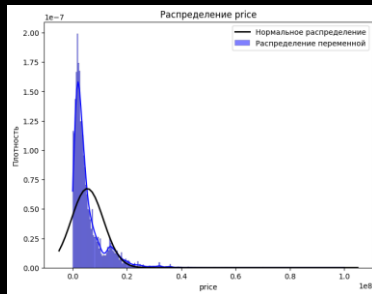
Распределение переменных

Тест Колмогорова-Смирнова

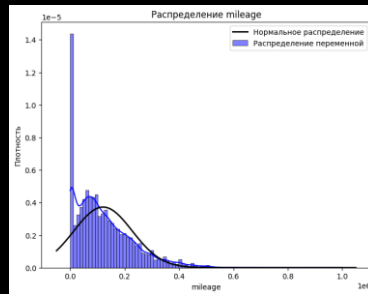
Исходя из показателей теста Колмогорова-Смирнова, у всех метрических переменных $p_value = 0$. Значит, принимаем гипотезу, что распределение сильно отличается от нормального.

Наблюдается:

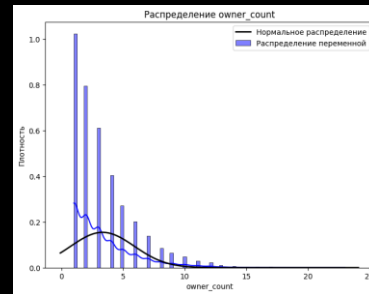
- сдвиг значений переменной `year` вправо
- сдвиг значений переменной `price` влево
- сдвиг значений числа владельцев влево
- сдвиг значений пробега влево



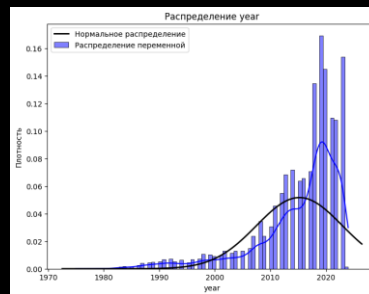
price



пробег



число владельцев



year

Корреляции

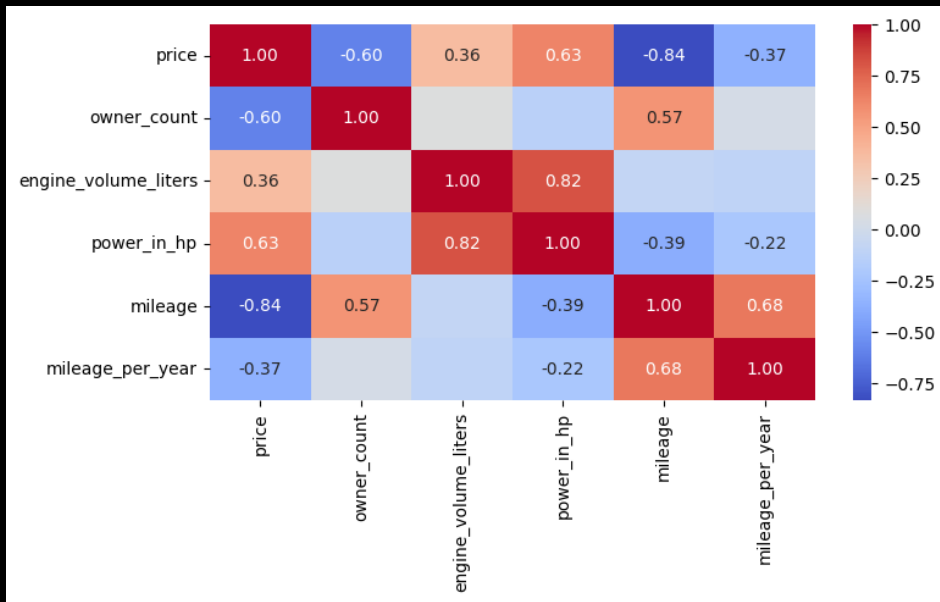
Коэффициент Спирмена:

- Сильная прямая между **объемом двигателя** и **числом лошадиных сил**
- Сильная обратная между **пробегом** и **ценой**
- Средняя прямая между **пробегом** и **пробегом за год**
- Средняя прямая между **ценой** и **числом лошадиных сил**
- Средняя прямая между **пробегом** и **количеством владельцев**
- Средняя обратная между **ценой** и **числом владельцев**
- Слабая прямая между **ценой** и **объемом двигателя**
- Слабая обратная между **пробегом** и **числом лошадиных сил**
- Слабая обратная между **ценой** и **пробегом за год**
- Слабая обратная между **пробегом за год** и **числом лошадиных сил**

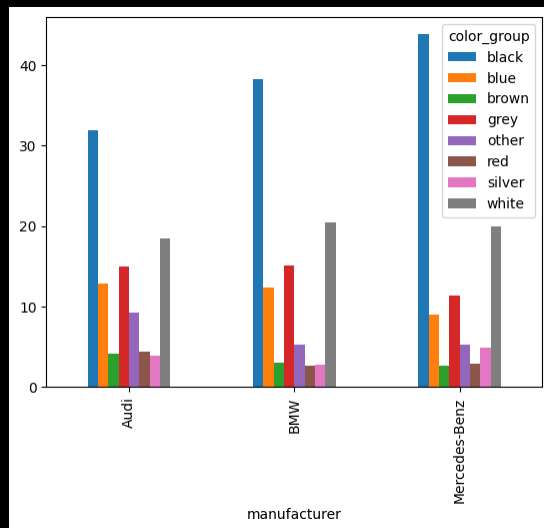


Корреляции

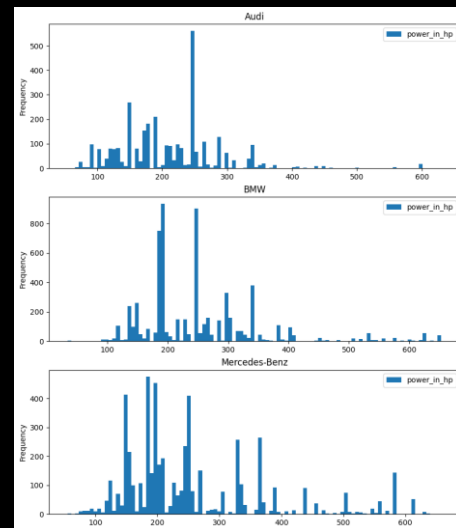
Коэффициент Спирмена:



Графики

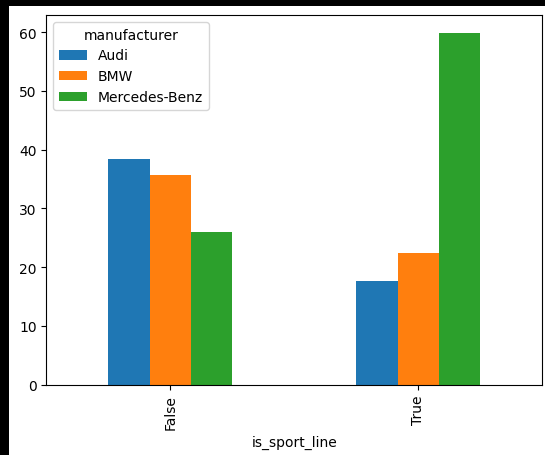


Цвет
по производителям

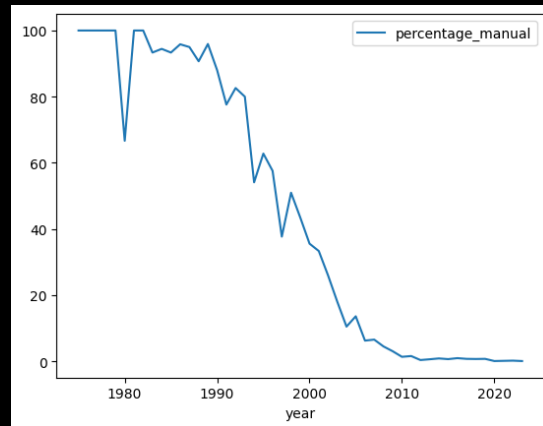


Мощность в л. с.
по производителям

Графики



Спортивная линейка
по производителям



Доля авто с МКПП
по годам

Тест Хи-Квадрат

Гипотезы:

1. Производитель и спортивный класс авто.

- H_0 : нет связи между производителем и принадлежностью к спортивному классу авто.
- H_1 : есть взаимосвязь между производителем и принадлежностью к спортивному классу авто.

$P\text{-value} < 0.05 \Rightarrow$ принимаем гипотезу H_1 .

2. Производитель и тип КПП.

- H_0 : нет связи между производителем и типом КПП.
- H_1 : есть взаимосвязь между производителем и типом КПП.

$P\text{-value} < 0.05 \Rightarrow$ принимаем гипотезу H_1 .



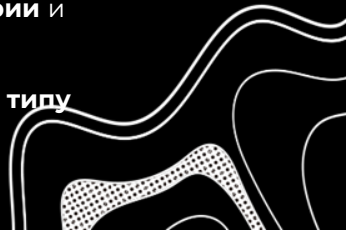
is_sport_line	False	True
manufacturer		
Audi	3133	81
BMW	2909	103
Mercedes-Benz	2126	275

manufacturer	Audi	BMW	Mercedes-Benz
transmission_group			
AT	2814	6044	5296
MT	400	278	197

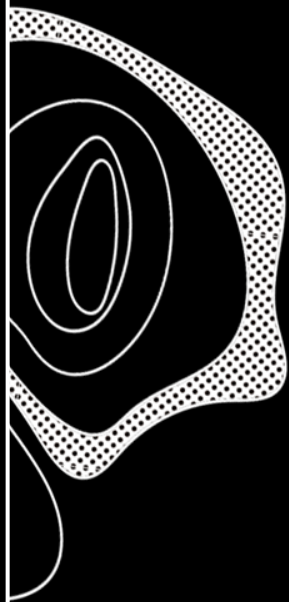
Тест Хи-Квадрат

Принятые гипотезы:

- H_1 : существует статистическая связь между **производителем** и **цветом**.
- H_1 : существует статистическая связь между **производителем** и **типом топлива**.
- H_1 : существует статистическая связь между **производителем** и **принадлежностью к кроссоверному типу**.
- H_1 : существует статистическая связь между **типом коробки передач** и **цветом**.
- H_1 : существует статистическая связь между **типом коробки передач** и **типом топлива**.
- H_1 : существует статистическая связь между **типом коробки передач** и **принадлежностью к спортивной серии**.
- H_1 : существует статистическая связь между **типом коробки передач** и **принадлежностью к кроссоверному типу**.
- H_1 : существует статистическая связь между **цветом** и **типом топлива**.
- H_1 : существует статистическая связь между **цветом** и **принадлежностью к спортивной серии**.
- H_1 : существует статистическая связь между **цветом** и **принадлежностью к кроссоверному типу**.
- H_1 : существует статистическая связь между **типом топлива** и **принадлежностью к спортивной серии**.
- H_1 : существует статистическая связь между **типом топлива** и **принадлежностью к кроссоверному типу**.
- H_1 : существует статистическая связь между **типом топлива** и **пробегом**.
- H_1 : существует статистическая связь между **принадлежностью к спортивной серии** и **принадлежностью к кроссоверному типу**.
- H_1 : существует статистическая связь между **принадлежностью к кроссоверному типу** и **пробегом**.



Кластеры



Индекс Калински-Харабаша

Максимальное расстояние при 3-х кластерах(≈ 4137) \Rightarrow берем 3 кластера.

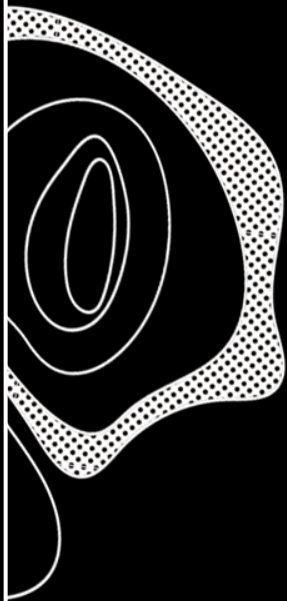
Кластер 1 - самые дорогие автомобили

- около 2017 года выпуска
- самые дорогие автомобили
- самые мощные, самый большой средний объем двигателя
- 13% - Audi, 42% - Mercedes, 45% - BMW
- 46% - черные, равномерное распределены (по 13-16%) синие, коричневые, белые и серые цвета
- В этом кластере доля черных наибольшая
- 8% автомобилей из спортивной линейки (самый высокий показатель)
- 64% автомобилей - кроссоверы (самый высокий показатель)
- Наиболее распространены премиальные модели (BMW X5, Mercedes G-class, Audi Q7)

Кластеры

Индекс Калински-Харабаша

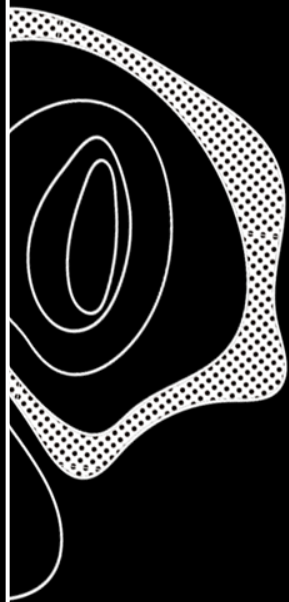
Максимальное расстояние при 3-х кластерах(≈ 4137) \Rightarrow берем 3 кластера.



Кластер 2 - "старички"

- около 1997 года выпуска
- самые дешевые
- самый большой пробег
- Равномерно распределены между BMW, Mercedes и Audi
- Наиболее распространенные цвета - черный, белый серебристый
- Меньше всего кроссоверов (11%)
- Больше всего механики (48%)
- Кластер содержит разнообразные модели

Кластеры



Индекс Калински-Харабаша

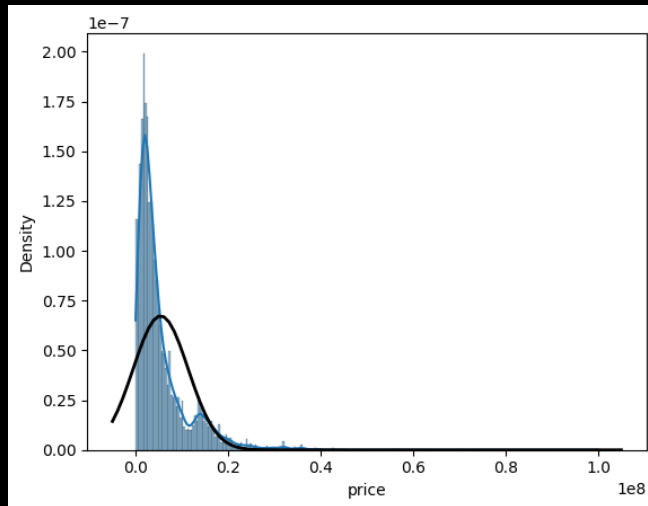
Максимальное расстояние при 3-х кластерах(≈ 4137) \Rightarrow берем 3 кластера.

Кластер 3 - "обычные"

- около 2015 года выпуска
- средние по цене
- Audi - 22%, равные доли у Mercedes и BMW
- Наибольшая доля белых автомобилей

Линейные регрессии

Предсказание цены
(переменная price)



1. Только изменяющиеся показатели

- 'mileage' – пробег;
- 'law_int' – есть ли проблемы на юридическом поле;
- 'owner_count' – количество владельцев;
- 'year' – год выпуска авто;
- 'is_new' – новый ли авто.

R² = 0.463996 (46%), **F-test significance** = 0.000000.

- **Mileage:** с увеличением пробега автомобиля на 1 километр, его цена снижается на 9.49 единиц.
- **Law Int:** Если есть проблемы с законом у автомобиля, его цена снижается на 1032612.66 единиц.
- **Owner Count:** С увеличением количества владельцев на 1, цена автомобиля снижается на 53450.11 единиц.
- **Year:** С увеличением года выпуска на 1, цена автомобиля повышается на 152932.73 единицы.
- **Is New:** Новизна авто связана с увеличением его цены на 8809467.64 единиц.

2. Добавим данные о двигателе, кузове, топливе

- 'is_diesel' – дизельный авто;
- 'sport_int' – спортивного ли класса авто;
- 'engine_volume_liters' – объем двигателя в литрах;
- 'power_in_hp' – мощность в л. с.;
- 'crossover_int' – автомобиль-кроссовер

R² = 0.587853 (58%), **F-test significance** = 0.000000.

- **Is Diesel:** тип двигателя не оказывает статистически значимого влияния на цену автомобиля.
- **Sport Int:** принадлежность авто к спортивному классу связана с увеличением цены автомобиля на 932976.45 единиц.
- **Engine Volume Liters:** С увеличением объема двигателя на 1 литр, цена автомобиля снижается на 610097.95 единиц.
- **Power in:** При увеличении мощности на 1 лошадиную силу, цена автомобиля повышается на 19253.33 единицы.
- **Crossover Int:** Принадлежность авто к кроссоверному типу связана с увеличением его цены на 792425.99 единиц.

3. Добавим информацию о КПП

- 'is_manual_transmission' – наличие механической КПП.
- ...
- ...
- ...
- ...

R² = 0.624347 (62%), **F-test significance** = 0.000000.

- **Is Manual Transmission (Механическая коробка передач):** Наличие механической коробки передач связано с увеличением цены на 3103190.13 единиц.
- ...
- ...
- ...
- ...
- ...
- ...
- ...

- 'manufacturer_Audi'* – производитель Audi;
- 'manufacturer_BMW' – производитель BMW;
- 'manufacturer_Mercedes-Benz' – производитель Mercedes
- ...
- ...

- **Manufacturer:** Для производителей (BMW: 646024.08, Mercedes-Benz: 938110.94) указаны соответствующие коэффициенты, по сравнению с референтной группой (Audi).

- [illegible]

5. Добавим информацию о цвете

- 'color_gr__blue_brown' – синий/коричневый;
- 'color_gr__grey_silver' – серый/серебряный
- 'color_gr__red' – красный;
- 'color_gr__white' – белый
- 'color_gr__other' – другие;

R2 = 0.637493 (63%), **F-test significance** = 0.000000.

- 'color_gr__blue_brown': Синий/коричневый дешевле черного на **35738992.04**.
- 'color_gr__grey_silver': Серый/серебряный дешевле черного на **35441814.23**. (незначим)
- 'color_gr__red': Красный дешевле черного на **35759599.05**.
- 'color_gr__white': Белый дешевле черного на **35803282.87**.
- 'color_gr__other': Другие цвета дешевле черного на **35445538.92**. (незначим)
- ...
- ...
- ...
- ...

Выводы:

На стоимость автомобиля влияют такие факторы как пробег, год выпуска, объем и мощность двигателя, принадлежность к кроссоверам или спортивной линейке или проблемы с законом. Модель статистически значимая (около 64%).

Неожиданными оказались следующие факторы:

- Отсутствие влияния типа двигателя;
- Меньший объем двигателя делает автомобиль дороже;
- Автомобили с механической коробкой дороже.



Бинарные регрессии

Бинарные регрессии

Предсказание Audi

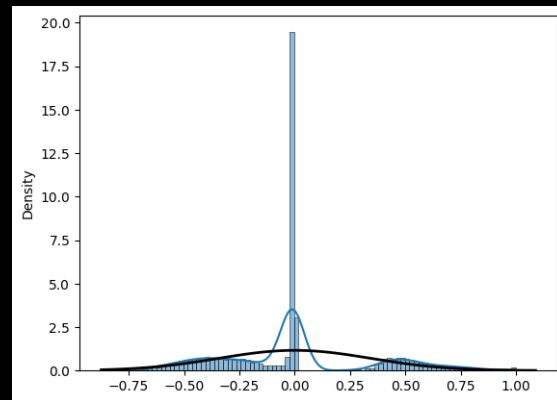
Модель статистически незначима, $R^2 = 32,2\%$.
Значимые предикторы: пробег, цена, объем двигателя, мощность, белый, серый цвет, коробка передач, принадлежность к спортивной линейке или линейке кроссоверов, число владельцев.

Features	Coefs
mileage	-0.194123
engine_volume_liters	-1.046885
power_in_hp	0.009507
color_gr_blue_brown	0.266334
color_gr_grey_silver	0.154562
color_gr_other	0.399820
color_gr_red	0.294855
color_gr_white	0.026782
is_manual_transmission	0.751156
year	-0.001295
price	-1.468057
sport_int	-4.227007
crossover_int	-0.190219
owner_count	-0.063672
law_int	-0.006224

Бинарные регрессии

Предсказание Audi

Распределение остатков отличается от нормального распределения, наблюдается сдвиг влево. Также есть проблема мультиколлинеарности цвета.



Распределение остатков

Бинарные регрессии

Предсказание Audi

Признаки того, что случайный автомобиль - Audi согласно данной модели:

- Меньший пробег
- Более низкая цена
- Более низкий объем двигателя
- Более высокая мощность
- Автомобиль не серый или не белый
- Механическая коробка передач
- Автомобиль не принадлежит к спортивной линейке
- Автомобиль - не кроссовер
- У автомобиля меньше предыдущих владельцев
- Более ранний год выпуска

Бинарные регрессии

Предсказание BMW

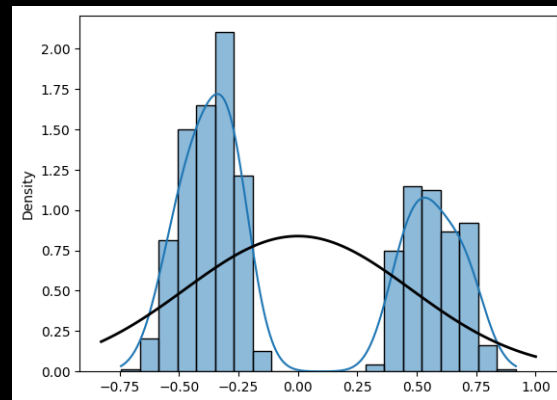
Модель статистически незначима, $R^2 = 4,05\%$.
Значимые предикторы: пробег, цена, объем двигателя, мощность, серый цвет, коробка передач, принадлежность к спортивной линейке или линейке кроссоверов, число владельцев, год выпуска, проблемы с законом, цена.

Features	Coefs
mileage	0.160378
engine_volume_liters	0.328203
power_in_hp	-0.002758
color_gr_blue_brown	0.103805
color_gr_grey_silver	0.175203
color_gr_other	-0.005337
color_gr_red	-0.192918
color_gr_white	0.080571
is_manual_transmission	0.405468
year	0.052582
price	-0.134195
sport_int	0.768309
crossover_int	0.403622
owner_count	0.024289
law_int	-0.229362

Бинарные регрессии

Предсказание BMW

Распределение остатков отличается от нормального распределения. Аналогично есть проблема мультиколлинеарности цвета.



Распределение остатков

Бинарные регрессии

Предсказание BMW

Признаки того, что случайный автомобиль - BMW согласно данной модели:

- Большой пробег
- Более низкая цена
- Большой объем двигателя
- Меньшая мощность
- Автомобиль серый или серебряный
- Механическая коробка передач
- Автомобиль принадлежит к спортивной линейке
- Автомобиль - кроссовер
- У автомобиля больше предыдущих владельцев
- Проблемы с законом

Бинарные регрессии

Предсказание Mercedes

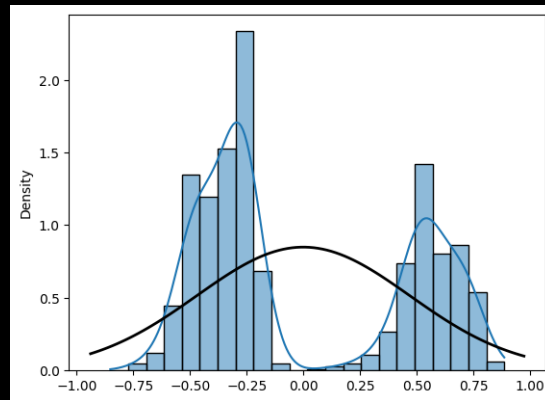
Данная модель не значима и описывает только 5% наблюдений. Значимыми переменными являются объем двигателя, мощность, коробка передач, год, цена, принадлежность к спортивным или кроссоверам автомобилям, и законодательный статус.

Features	Coefs
mileage	-0.045608
engine_volume_liters	0.254897
power_in_hp	-0.002185
color_gr_blue_brown	-0.264205
color_gr_grey_silver	-0.259319
color_gr_other	-0.255316
color_gr_red	-0.026218
color_gr_white	-0.104089
is_manual_transmission	-1.054007
year	-0.037444
price	0.553394
sport_int	0.856049
crossover_int	-0.258632
owner_count	0.011877
law_int	0.215811

Бинарные регрессии

Предсказание Mercedes

Распределение остатков отличается от нормального распределения. Аналогично есть проблема мультиколлинеарности цвета.



Распределение остатков

Бинарные регрессии

Предсказание Mercedes

Среднестатистический автомобиль марки Mercedes:

- Более высокий объем двигателя
- Более низкая мощность
- Автоматическая коробка передач
- Более поздний год выпуска
- Более высокая цена
- Спортивная линейка
- Больше число владельцев
- Проблемы с законом

Спасибо за внимание!

1

Три кластера авто:

Дорогие
Старые
Обычные

2

Цена зависит от:

- + мощность
- + год выпуска
- + кроссовер
- + МКПП
- + спорт
- объем двигателя
- пробег
- незаконность

3

Предсказания*:

МКПП => Audi
Кроссовер => BMW
Спорт => Mercedes

*статистически незначимы