



데이터톤 4조

강경희, 나상현, 박서윤, 이상원

CUAI

DArt-B

Get Started >>

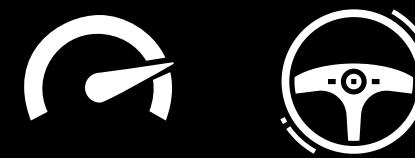
Home

About

Contact



목 차



1.EDA

2.전처리

3.모델링

4.인사이트



Home

About

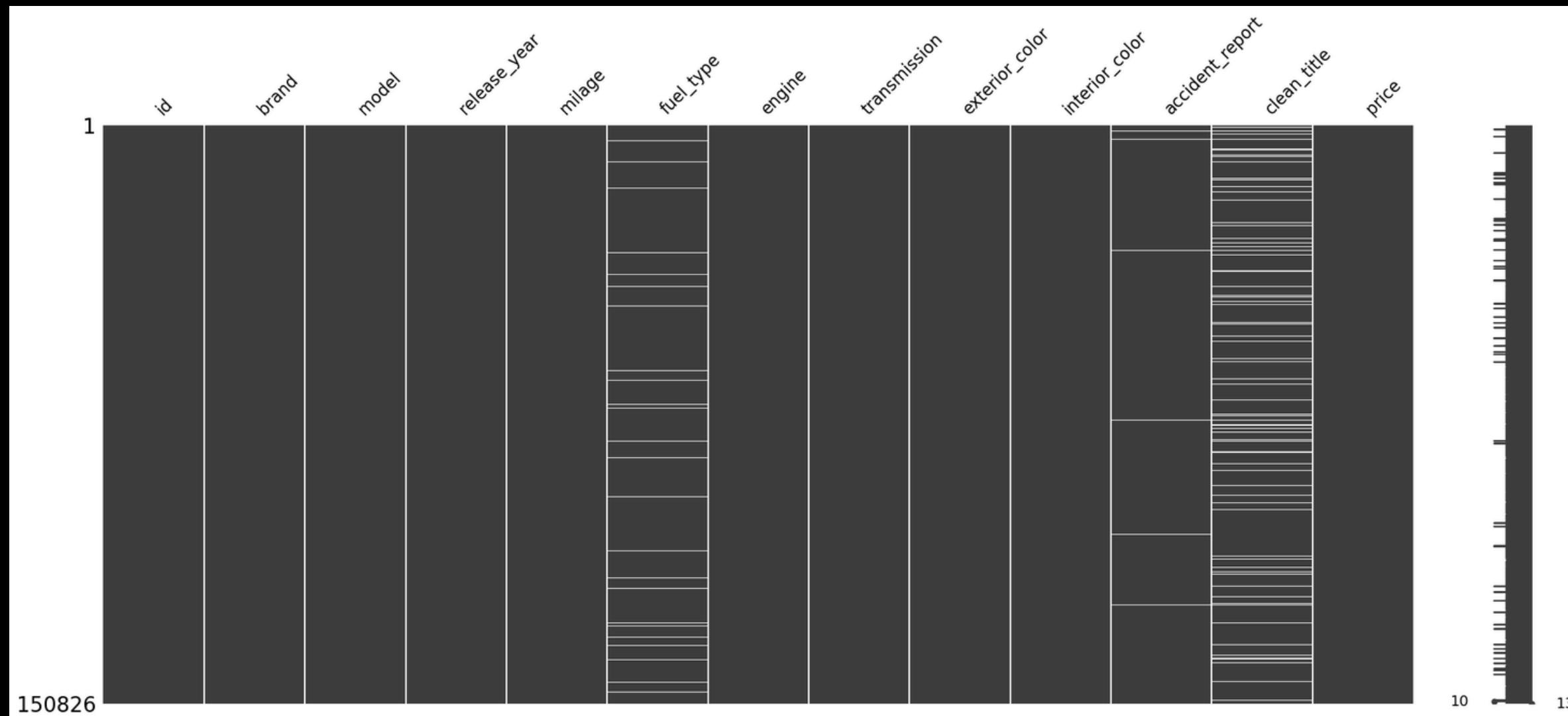
Contact



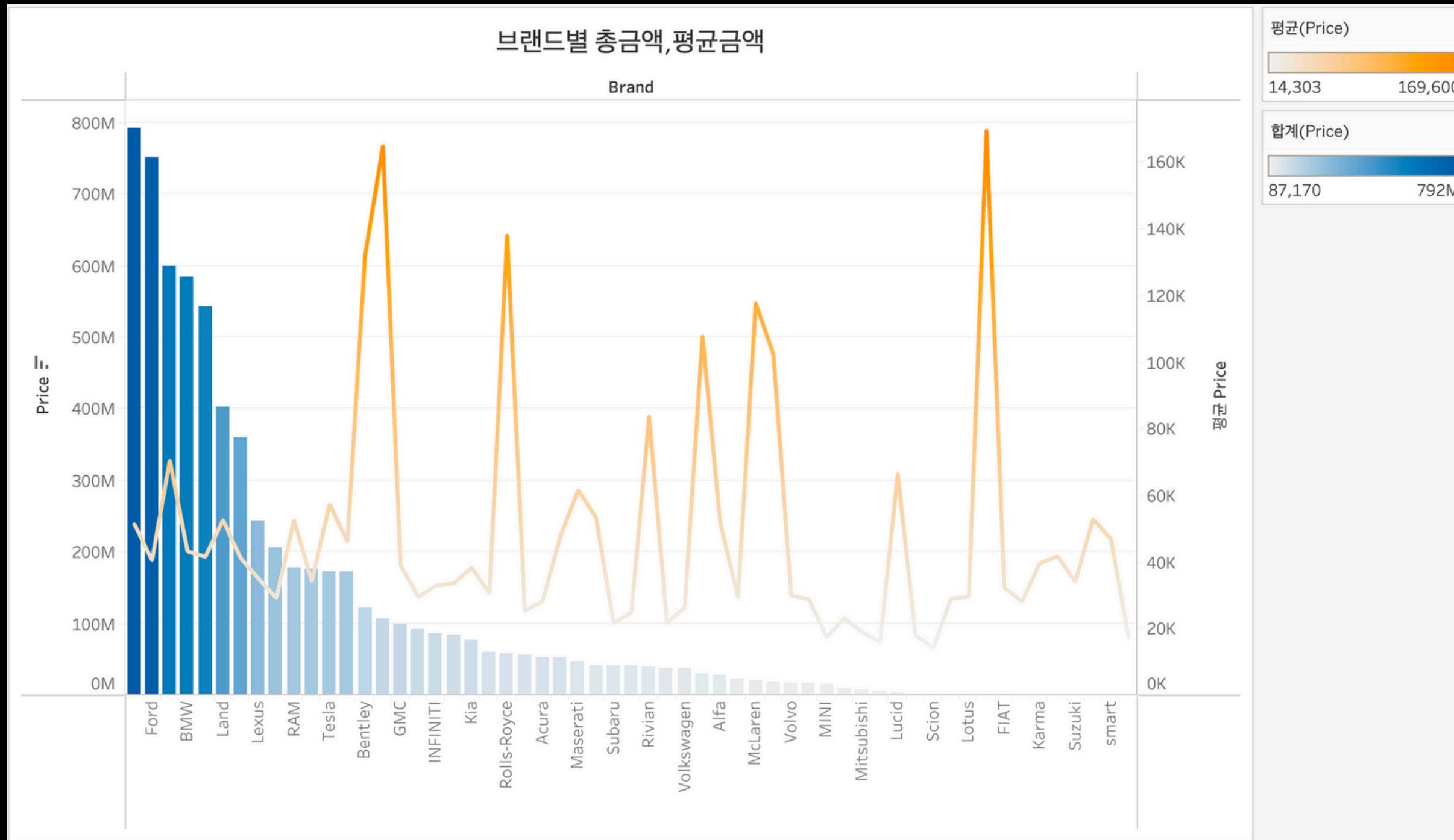
EDA

Learn More

결측치 확인



| | 결측치 수 | 결측치 비율 (%) |
|-----------------|-------|------------|
| id | 0 | 0.00 |
| brand | 0 | 0.00 |
| model | 0 | 0.00 |
| release_year | 0 | 0.00 |
| milage | 0 | 0.00 |
| fuel_type | 4085 | 2.71 |
| engine | 0 | 0.00 |
| transmission | 0 | 0.00 |
| exterior_color | 0 | 0.00 |
| interior_color | 0 | 0.00 |
| accident_report | 1991 | 1.32 |
| clean_title | 17205 | 11.41 |
| price | 0 | 0.00 |

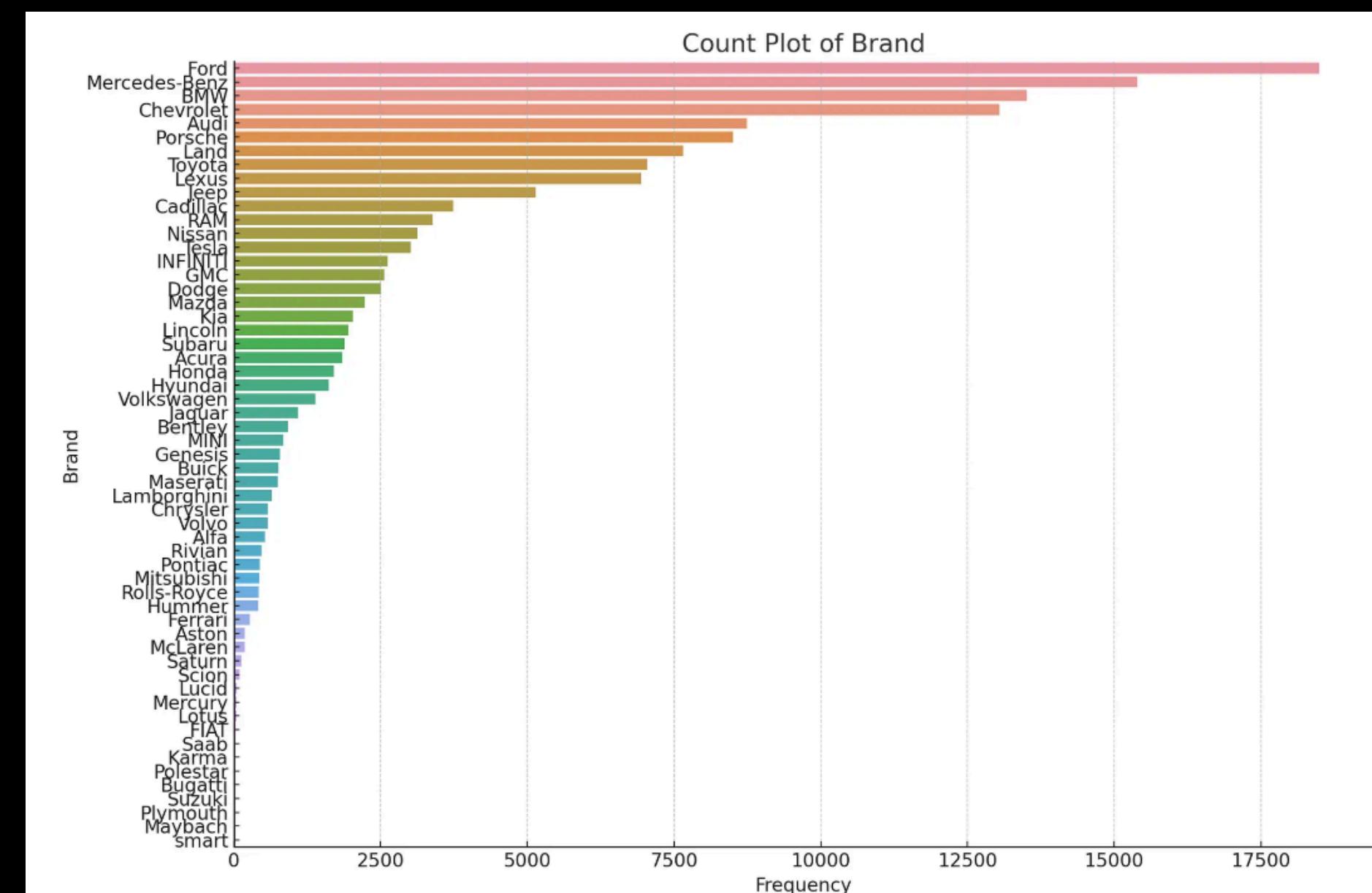


test 데이터셋 브랜드별 총금액 순위

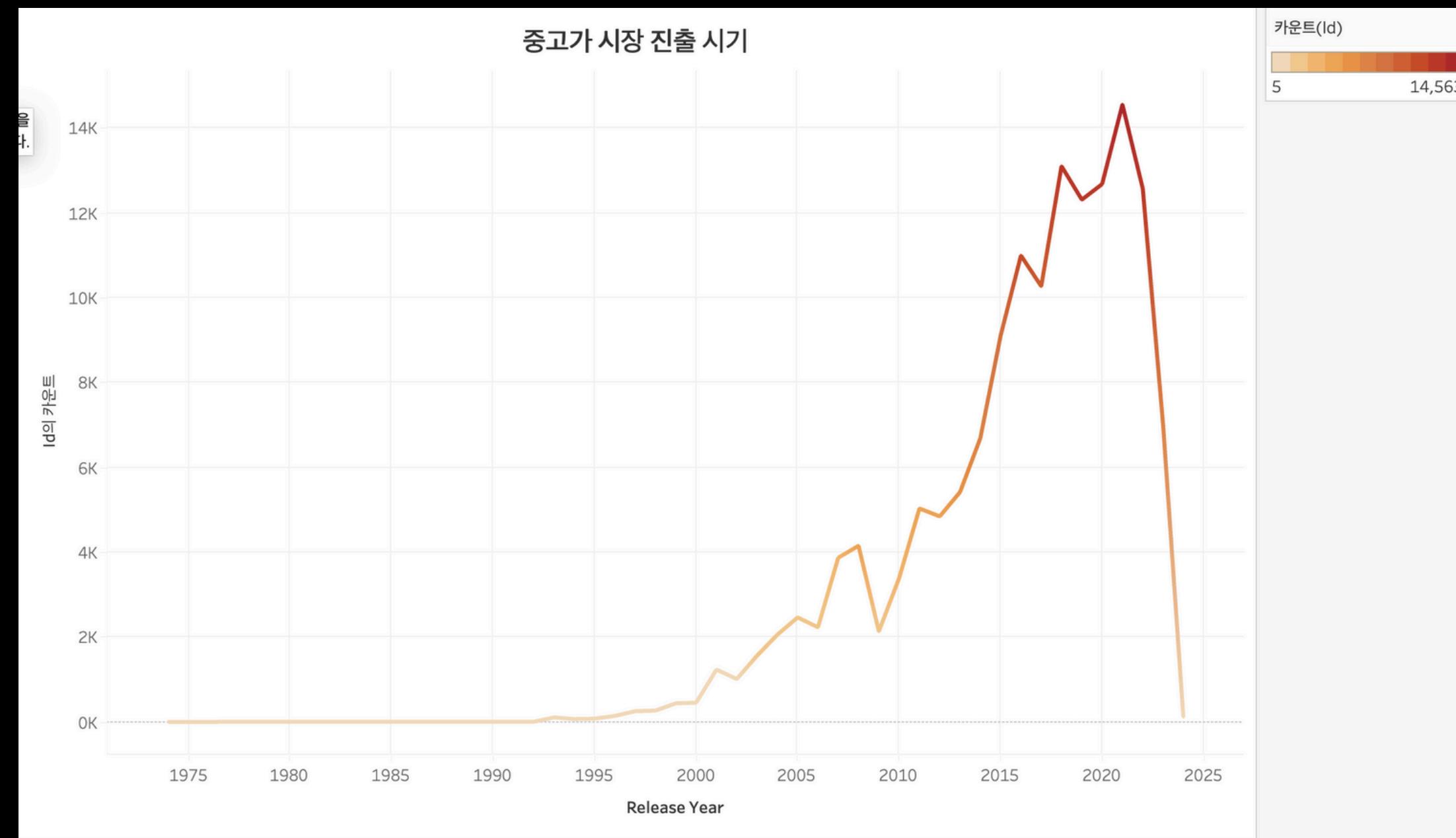
1. 메르세데스벤츠
2. 포드
3. 포르쉐
4. BMW
5. 쉐보레

브랜드별 평균금액 기준

1. 부가티
2. 람보르기니
3. 롤스로이스
4. 맥라렌
5. 페라리

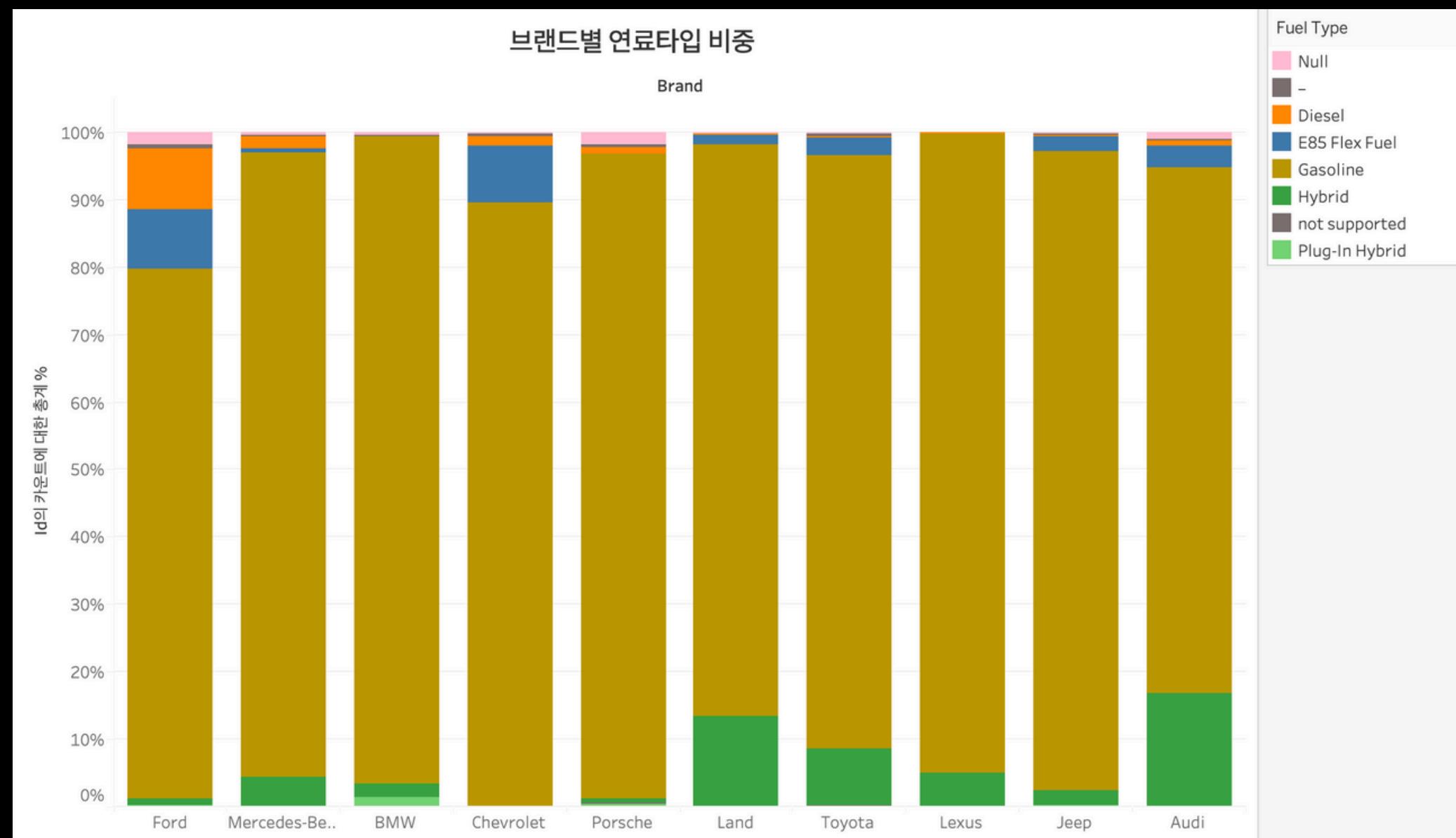


브랜드 별 중고차 수 분포



2021년, 2018년, 2020년, 2022년 순으로 출고시기가 가장
많고, 대부분의 차량이 최근 15년 이내에 출고되었음

연료 타입 관련



가솔린 타입이 가장 많음

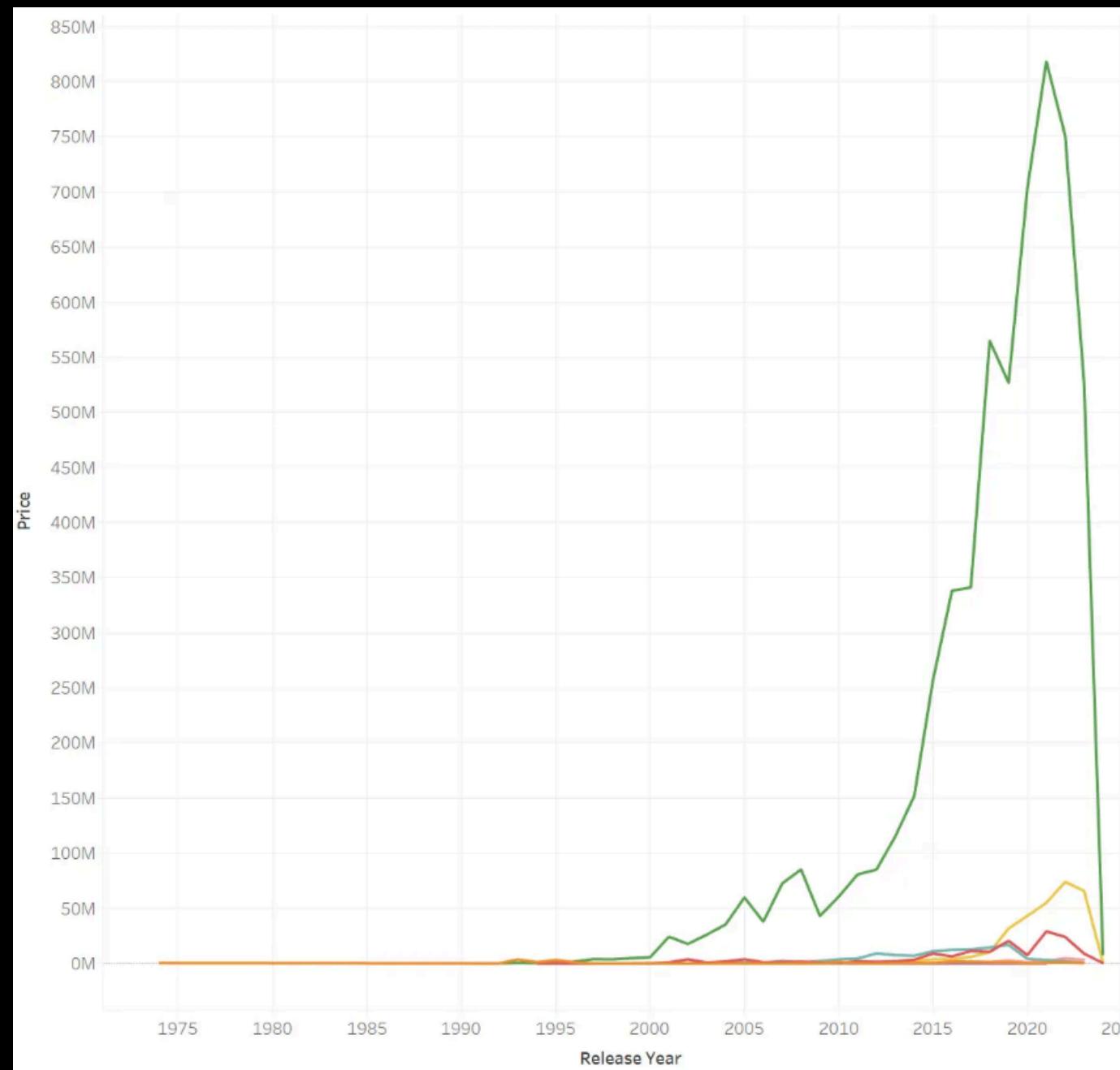
E85 flex fuel 타입

미국의 넓은 대지 지형학적 특성상 가솔린이 대다수이지만,
포드, 쉐보레 등에서 에탄올과 가솔린이 결합한 친환경적인
E85 flex fuel 타입 비율이 있음

하이브리드 타입

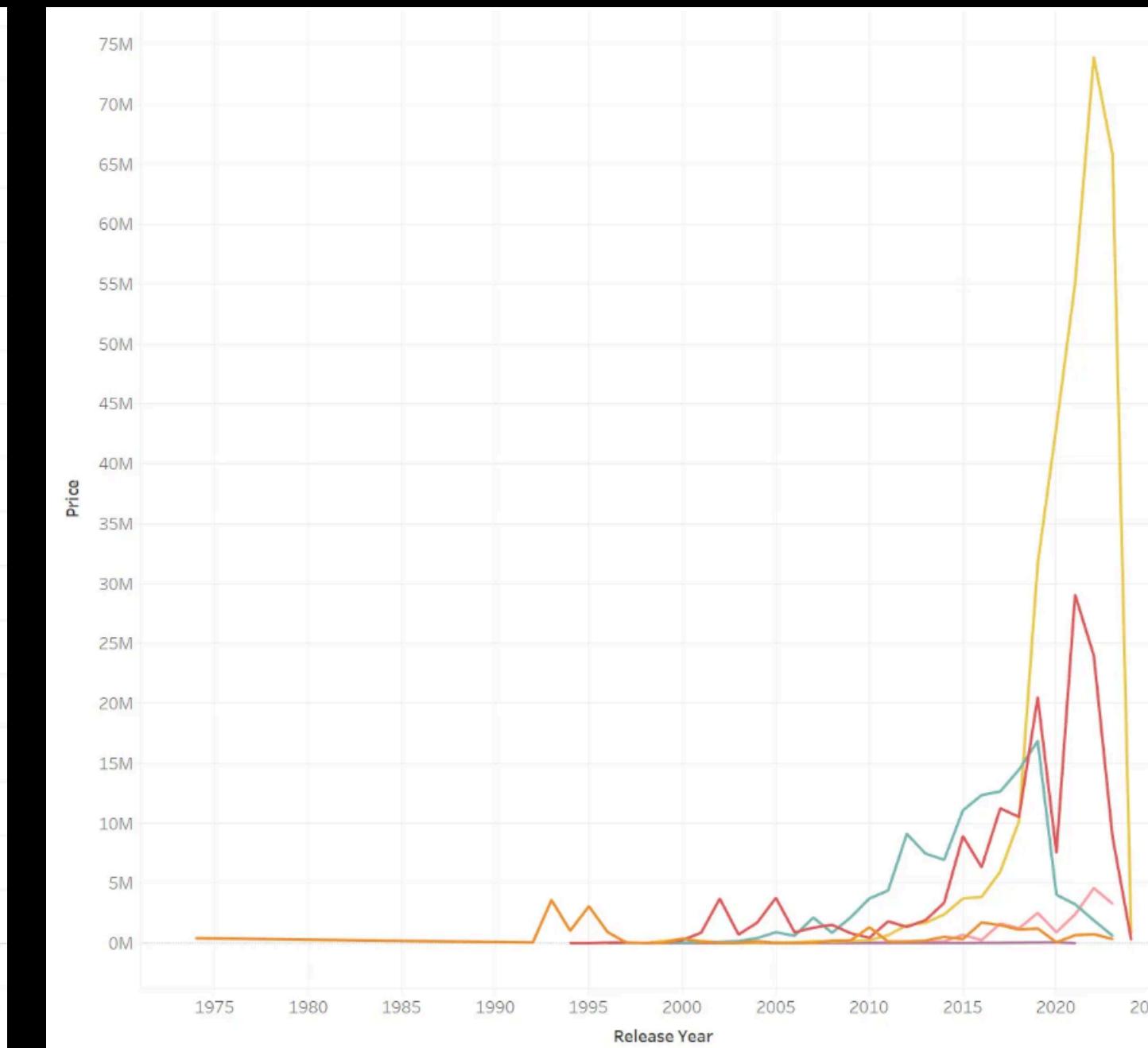
아우디, 랜드로버의 경우 하이브리드 차량 비율이 타 브랜드에 비해 많음

'fuel type'별로 연도별 가격 분포



1. 'Gasoline'

fuel type이 'gasoline', 'hybrid'인 차량의 경우 중고차 시장에서의 수요가 높다.



2. 'Hybrid'

| Fuel Type |
|----------------|
| - |
| Diesel |
| E85 Flex Fuel |
| Gasoline |
| Hybrid |
| not supported |
| Plug-In Hybrid |

Home

About

Contact



전처리

Learn More

PREPROCESS



```
▶ data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150826 entries, 0 to 150825
Data columns (total 13 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   id          150826 non-null  int64  
 1   brand        150826 non-null  object 
 2   model        150826 non-null  object 
 3   release_year 150826 non-null  int64  
 4   milage       150826 non-null  object 
 5   fuel_type    146741 non-null  object 
 6   engine        150826 non-null  object 
 7   transmission 150826 non-null  object 
 8   exterior_color 150826 non-null  object 
 9   interior_color 150826 non-null  object 
 10  accident_report 148835 non-null  object 
 11  clean_title   133621 non-null  object 
 12  price         150826 non-null  int64  
dtypes: int64(3), object(10)
memory usage: 15.0+ MB
```

release_year: int
milage:
60.0k=> 60.0(float)
clean_title:
Yes=>1
null=>0
accident_report:
1 or more accident reported=>1
None reported =>0
brand:
model과의 다중공산성 issue로 제거

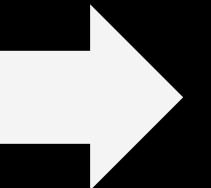
PREPROCESS

Transmission Column



```
integ['transmission'].unique()
```

```
array(['Automatic Transmission', 'Transmission with Dual Shift Mode', '6-Speed Automatic Transmission', '7-Speed Automatic Transmission', 'Automatic', '9-Speed Automatic', 'Transmission Overdrive Switch', '6-Speed Manual Transmission', 'CVT Transmission', '8-Speed Automatic Transmission', '10-Speed Automatic', '8-Speed Automatic', '10-Speed Automatic Transmission', '4-Speed Automatic Transmission', '6-Speed Manual', '9-Speed Automatic Transmission', '5-Speed Manual Transmission', '10-Speed Automatic with Overdrive', '1-Speed Automatic Transmission', '6-Speed Automatic', '7-Speed Automatic with Auto-Shift', 'Manual Transmission', '5-Speed Automatic Transmission', '7-Speed Manual Transmission', '7-Speed Automatic', 'Automatic CVT', '8-Speed Automatic with Auto-Shift', '7-Speed', 'F', '6-Speed', '1-Speed Automatic', '2', '7-Speed Manual', 'Variable', '—', '4-Speed Automatic', '5-Speed Automatic', '6-Speed Automatic with Auto-Shift', '2-Speed Automatic Transmission', 'CVT-F', '8-SPEED AT', '6 Speed Manual Transmission', '6-Speed Electronically Controlled Automatic with O', '9-Speed Automatic with Auto-Shift', '8-Speed Manual', '2-Speed Automatic', '8-SPEED Automatic Transmission', 'Manual', 'Single-Speed Fixed Gear', 'SCHEDULED FOR OR IN PRODUCTION', '7-Speed DCT Automatic', '6 Speed At/Manual Transmission'],  
dtype=object)
```



| simplified_transmission | count |
|--|-------|
| Automatic | 60751 |
| 8-Speed Automatic | 29096 |
| 6-Speed Automatic | 20874 |
| Transmission with Dual Shift Mode | 19255 |
| 6-Speed Manual | 12246 |
| 7-Speed Automatic | 11302 |
| 10-Speed Automatic | 11208 |
| 9-Speed Automatic | 6191 |
| 5-Speed Automatic | 3376 |
| CVT | 2900 |
| 4-Speed Automatic | 2594 |
| 5-Speed Manual | 2409 |
| 1-Speed Automatic | 2250 |
| Manual | 1206 |
| 7-Speed Automatic with Dual Shift Mode | 1172 |
| 8-Speed Automatic with Dual Shift Mode | 549 |
| 7-Speed Manual | 506 |
| Other | 249 |
| 6-Speed Automatic with Dual Shift Mode | 111 |
| 2-Speed Automatic | 93 |
| Automatic with Overdrive | 40 |
| 9-Speed Automatic with Dual Shift Mode | 33 |
| 7-Speed Transmission | 29 |
| 8-Speed Manual | 28 |
| 6-Speed Transmission | 27 |
| 6-Speed Automatic with Manual Mode | 20 |
| 7-Speed DCT Automatic | 18 |

CVT, Automatic, Manual의
3가지로 범주화 + 숫자별 변속기로 분류

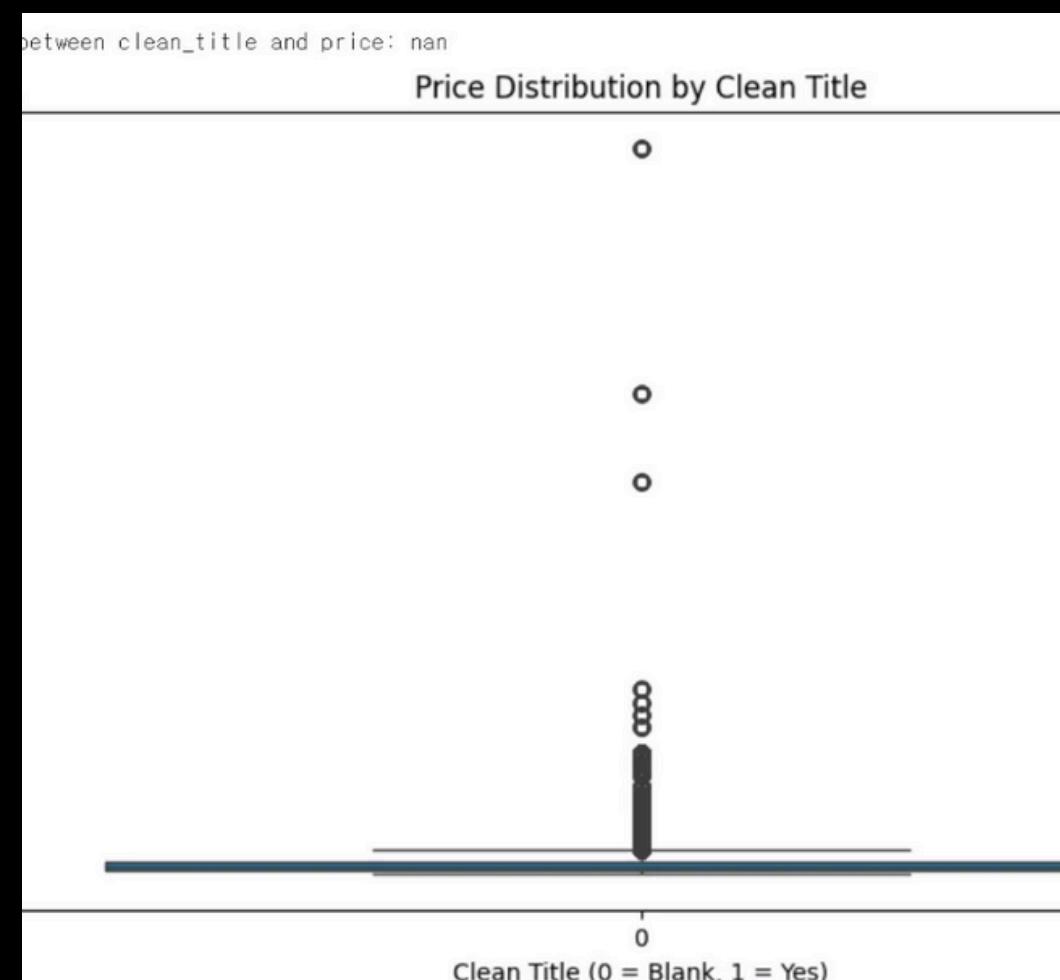
지저분하게 되어 있는 index들 정리

ex.

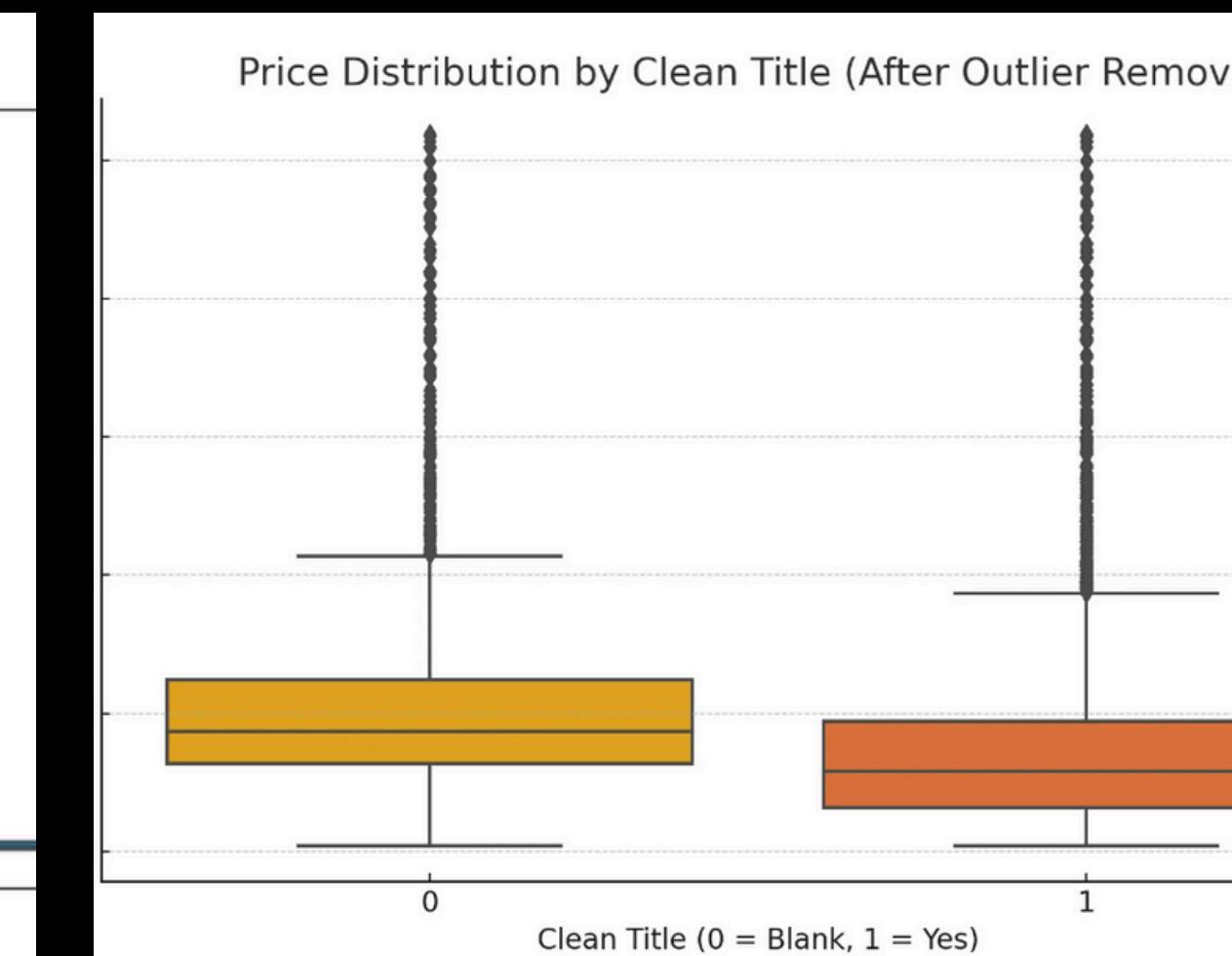
Variable → CVT

Transmission with Overdrive → Automatic
Fixed Gear → C1-Speed Automatic

'clean title' 여부에 따른 가격 분포 (box plot)



이상치 제거 전



이상치 제거 후



clean_title와 price 간의 상관계수 = 약 -0.167

'독립 표본 t-검정'

Levene's Test

Levene 통계량: 211.43 귀무가설을 기각
p-value: 7.22×10^{-48}

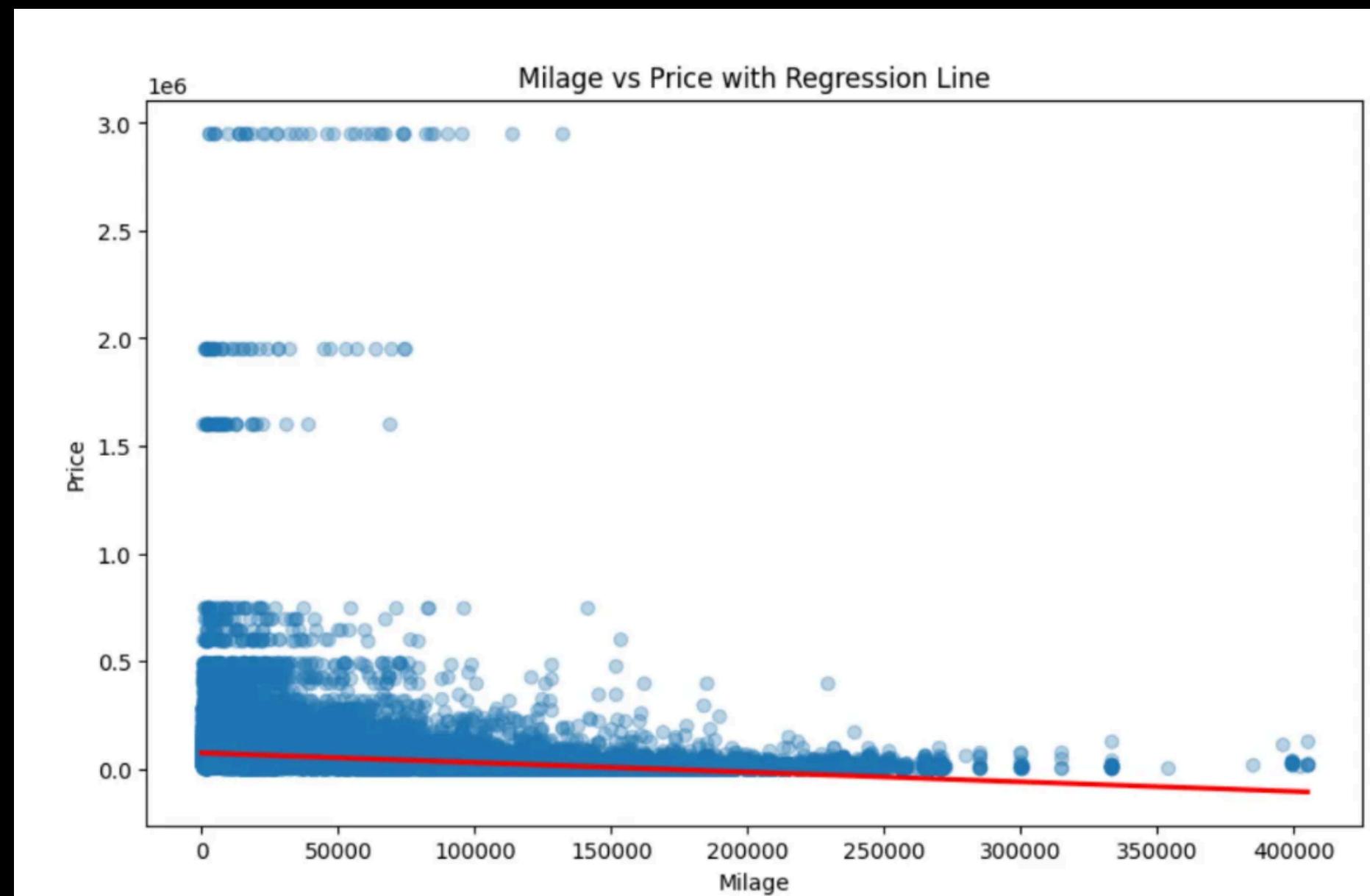
두 그룹(clean_title_binary = 0과 clean_title_binary = 1)의 분산이 같지 않다

-> Welch's t-test

t-통계량: 55.31 귀무가설을 기각
p-value: 0.0

두 그룹('clean_title'이 '0'인 그룹과 '1'인 그룹) 간 평균 price에 차이가 없다

'clean_title'이 'price'에 유의미한 영향을 준다



```
pearson_corr, pearson_p_value  
(-0.2858281710368217, 0.0)
```

피어슨 상관계수가 -0.286 → 약한 음의 상관관계를 가진다.
또한, $p\text{-value}=0 < 0.05$ → **상관관계가 유의미**하다.

'milage'와 'price' 간의 산점도를 회귀선과 함께 나타냈음.

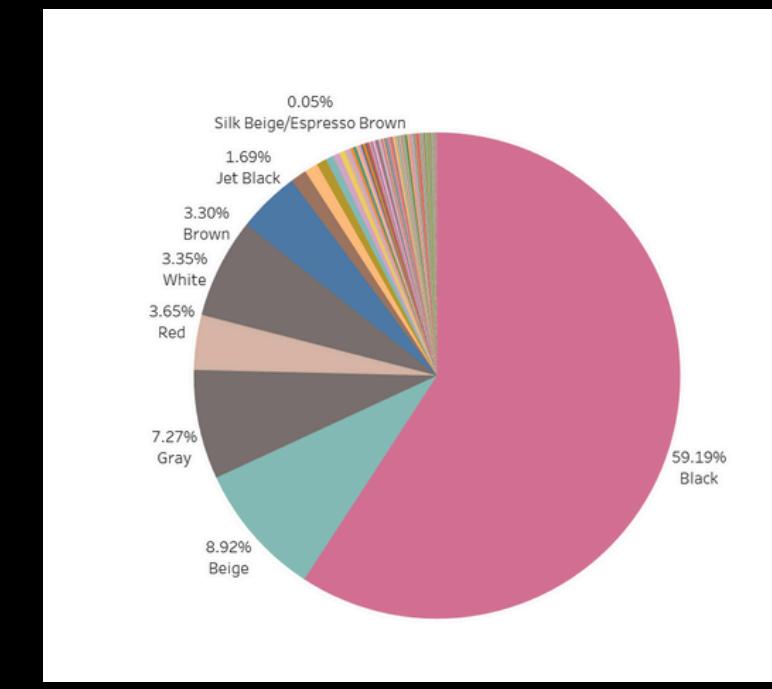
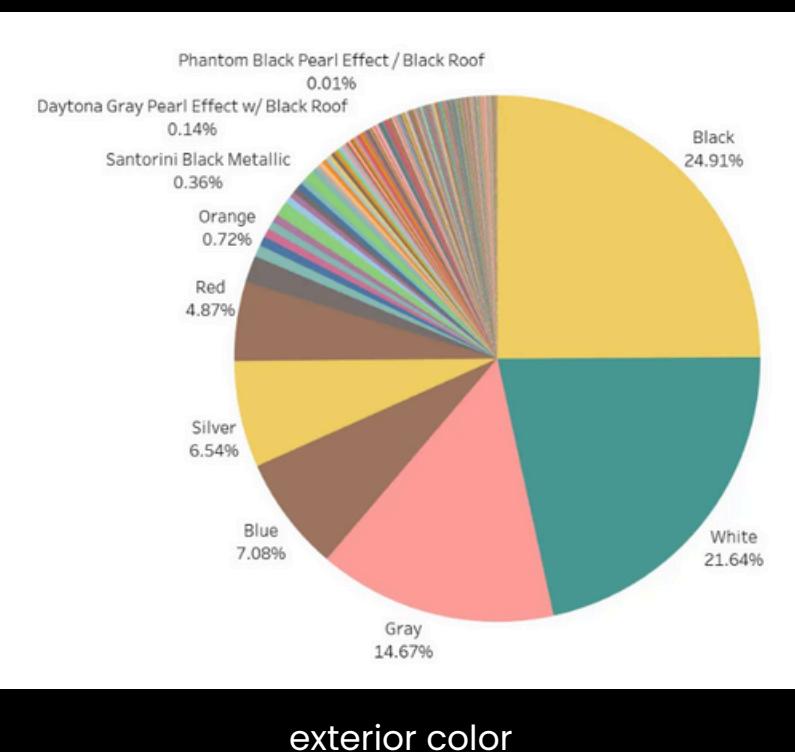
PREPROCESS

interior_color & exterior_color Columns



```
Value counts for exterior_color:  
exterior_color  
Black      38984  
White     35076  
Gray      20333  
Silver    13541  
Blue      11569  
...  
Blue Caelum    4  
GT SILVER      4  
Pure White     3  
BLUE          3  
Mango Tango Pearlcoat 1  
Name: count, Length: 319, dtype: int64
```

```
Value counts for interior_color:  
interior_color  
Black     86118  
Beige    19580  
Gray     16994  
Brown    4634  
Red      4091  
...  
Cobalt Blue   6  
WHITE        5  
Classic Red   5  
Bianco Polar  4  
ORANGE        4  
Name: count, Length: 156, dtype: int64
```



interior 상위 6개+비주류, exterior 상위 4개+비주류
로 조정 후 이들 경우의 수로 조합한 칼럼 생성

미국 중고차 시장에서 무채색(화이트, 블랙, 그레이, 실버)의 인
기가 지속되고 있으며, 전체 시장의 약 80%를 차지

출처:

<https://www.repairerdrivennews.com/2024/06/24/grayscale-colors-continue-to-hold-market-majority-of-used-car-sales/>

PREPROCESS

interior_color & exterior_color Columns



페이지 iii 열 Exterior Color
행 행 Brand F

필터 Brand Exterior Color

마크 자동 색상 크기 텍스트 세부 정보 도구 설명 Exterior Color 합계(Price)

시트 12

| Brand | Exterior Color | | | | | |
|-------------|----------------|---------------|-------------|-----------|---------------|-----------|
| | Beluga Black | Bianco Icarus | Blue Caelum | C / C | Dark Sapphire | GT SILVER |
| Bentley | 10,882,566 | | 3,742,685 | 3,623,523 | | 6,930,093 |
| Porsche | | 279,900 | 1,950,995 | 837,436 | 209,995 | |
| Rolls-Royce | | | 234,900 | 944,915 | 1,759,825 | |
| McLaren | | 187,900 | | | | |
| Lamborghini | 13,041,769 | 1,599,000 | | | | |
| Ferrari | 274,900 | | | | | |
| Aston | 259,000 | | 399,950 | | | |

Exterior Color
 (전체)
 -
 Agate Black Metalic...
 Alfa White
 Alpine White
 Alta White
 Ametrin Metallic...
 Anodized Blue M...
 Anthracite Blue ...
 Antimatter Blue ...
 Apex Blue
 Arancio Borealis
 Arctic Gray Metalic...
 Arctic White
 Atomic Silver
 Aurora Black
 Aventurine Green...
 Balloon White
 Baltic Gray
 Barcelona Red
 Bayside Blue
 Beige
 Beluga Black
 Bianco Icarus Metalic...
 Bianco Isis
 Bianco Monocerus
 Billet Clearcoat...
 Billet Silver Metalic...
 Black
 BLACK
 Black Cherry

제한
 상위 10개(AVG([Price])기준)

Exterior Color
■ Beluga Black
■ Bianco Icarus Metalic...
■ Blue Caelum
■ C / C
■ Dark Sapphire
■ GT SILVER
■ Ice
■ Tempest

평균 가격 상위 10개의 차를 brand 별로 구분

exterior color, 해당되는 평균 가격 시각화

'beluga black', 'bianco icarus metallic'

-> 각각 'black', 'white' 색상

-> 앞에서 'black' 색의 중고차의 수요가 가장 많았음을 뒷받침.



250.0HP 3.5L V6 Cylinder Engine Gasoline Fuel

4.0L V8 32V GDI DOHC Twin Turbo

2.0 Liter Supercharged

...

엔진 피처에는 위와 같이 여러 칼럼으로 분리될 수 있는 내용들이 포함되었음
(엔진 스펙 작성 방식도 모두 상이하였음)

모델이 이를 쉽게 파악하도록 하기 위해 여러 칼럼으로 분리하는 전처리를 수행함



Home

About

Contact



모델링

Learn More



사용한 모델 : AutoGluon

AutoGluon은 아마존에서 개발한 오픈소스 AutoML 툴킷으로, 간단한 코드만으로도 좋은 모델을 구현할 수 있게 해주는 라이브러리입니다.

기본적인 전처리를 수행 후,
여러 모델(RF, KNR, LGBM, XGBoost, CatBoost etc.)을 학습하고
다층 스태킹 앙상블하여 결과를 제공합니다.

사용한 이유 :

Feature Engineering과 도메인 지식 공부에 시간을 더 투자하기 위해서
여러 모델을 학습시키고 제일 성과가 좋았던 모델을 사용하는 시간을 줄이기 위해서

Home

About

Contact



인사이트 도출

Learn More



가격에 큰 영향을 미친 *FEATURE*

| Feature | Importance | Stddev |
|-------------------------|-------------|------------|
| milage | 4618.440236 | 433.718848 |
| model | 2095.517932 | 999.860522 |
| release_year | 1018.858335 | 242.953849 |
| engine | 966.051515 | 691.796802 |
| brand | 894.765142 | 563.587842 |
| HorsePower | 681.940973 | 192.359328 |
| color_combination | 389.353473 | 247.977373 |
| simplified_transmission | 253.283415 | 204.098323 |
| accident_report | 172.774689 | 204.966418 |
| Cylinder | 153.319210 | 74.143368 |
| Liter | 70.930186 | 39.394776 |
| fuel_type | 4.031235 | 4.151354 |



주행 거리(milage):

미국은 넓은 국토를 가진 나라이므로 도시 간의 거리가 멀고, 장거리 운전이 흔함. 따라서 주행 거리가 길어질수록 유지 보수 비용이 증가할 가능성등으로 차량의 가치가 크게 떨어지는 경향이 있음.

모델(model): 특정 모델에 대한 수요와 브랜드 인식, 유지비, 연비, 안전성 등이 중고차 가격에 영향을 미침. 이는 미국의 일부 지역에서는 오프로드 또는 장거리 주행에 적합한 차량 모델이 더 인기가 많다는 것을 반영

출시 연도(release_year): 최신 모델일수록 더 비싼 경향이 있으며, 전기차나 하이브리드 모델의 경우 최신 연식 차량의 수요가 증가하고음

엔진(engine)과 **마력(HorsePower)**: 미국은 큰 차량과 강력한 엔진을 선호하는 시장이기 때문에, 엔진 성능이 좋은 차량이 중고차 시장에서 높은 가격을 받을 가능성

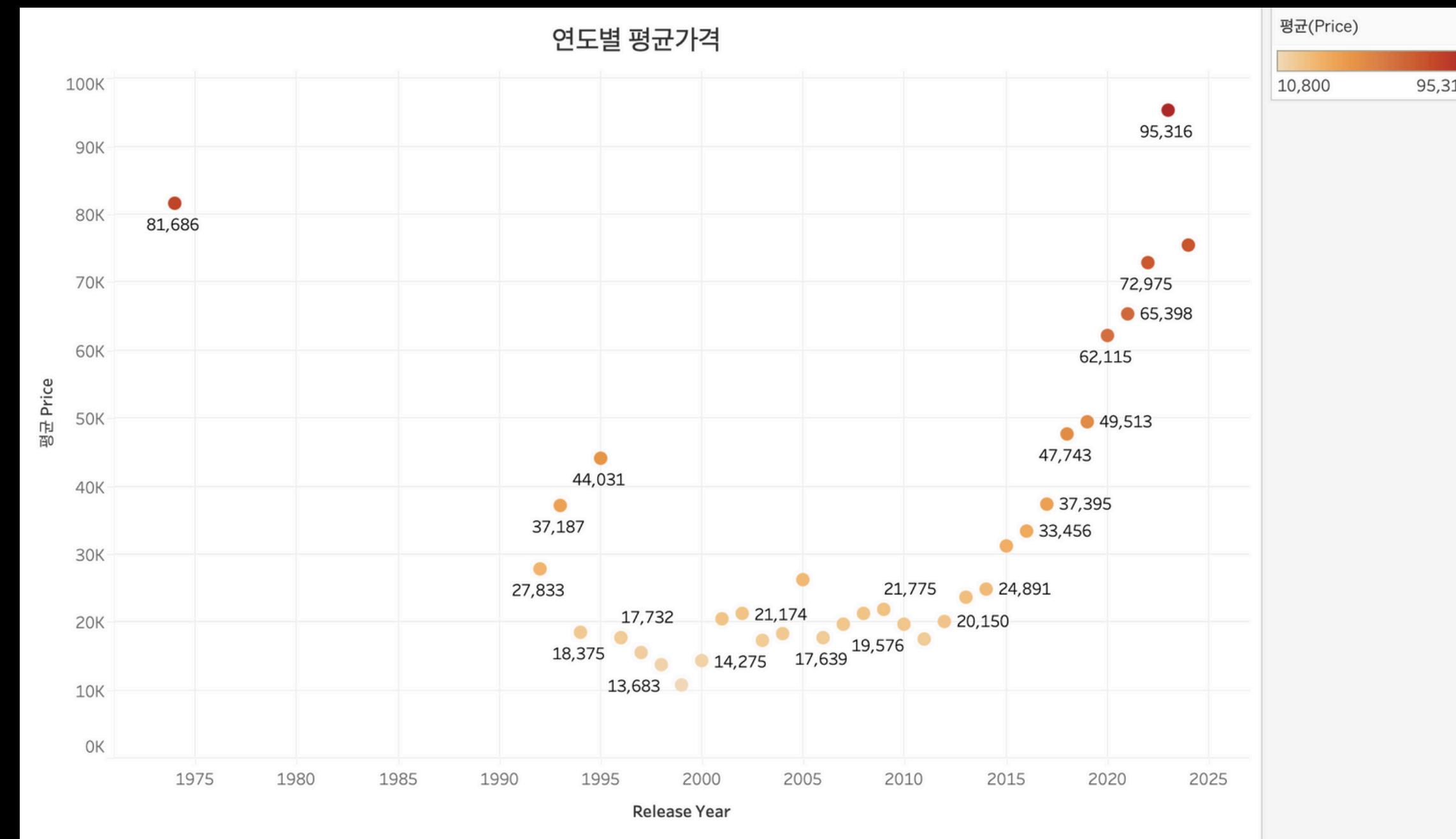
브랜드(brand): 고급 브랜드(예: BMW, Mercedes-Benz)는 중고차 시장에서도 높은 가치를 가지며, 신뢰성과 브랜드 이미지가 높은 브랜드는 중고차에서도 높은 가격을 유지할 가능성이 큼

색상 조합(color_combination): 선호와 비선호 색상으로 나뉘며. 특정 색상 조합은 중고차 가격에 영향을 미치는 임

변속기 유형(simplified_transmission): 대부분의 미국 소비자는 자동 변속기를 선호하므로, 자동 변속기 차량이 상대적으로 높은 가치를 가지는 경향이 있을 수 있습니다.

사고 이력(accident_report): 고 차량은 소비자들이 수리 비용과 차량의 안전성에 대한 우려로 인해 중고차 시장에서 낮은 가격으로 거래되는 경향이 있습니다. 이로 인해 사고 이력 없는 차량은 더 높은 가치를 가질 가능성이 있습니다.

레트로 차량의 인기



특정적으로, 항상 출고연도가 가까울수록 가격이 높아지는 것이 아님
1990년대에 출시된 “레트로카”인 경우 평균 가격이 높은 경향이 있음

Home

About

Contact



Q & A

4조

