DATA PREPROCESSING

1. feature selection

missing value ratio

각 feature의 결측치 비율 (%):

vehicle_damage_category	100.000000
combine_fuel_economy	100.000000
is_certified	100.000000
bed	99.347742
cabin	97.882262
is_oemcpo	95.487993
is_cpo	93.903481
bed_height	85.696924
bed_length	85.696924
owner_count	50.566426
fleet	47.552533
theft_title	47.552533
isCab	47.552533
has_accidents	47.552533
frame_damaged	47.552533
salvage	47.552533
franchise_make	19.087579
torque	17.259537
highway_fuel_economy	16.375948
city_fuel_economy	16.375948
power	16.047319
interior_color	12.799363
main_picture_url	12.302936
major_options	6.668178
engine_displacement	5.746123
horsepower	5.746123
back_legroom	5.308896
wheelbase	5.308896
maximum_seating	5.308896
width	5.308896
length	5.308896
height	5.308896
front_legroom	5.308896
fuel_tank_volume	5.308896
wheel_system_display	4.891001
wheel_system	4.891001
mileage	4.812836
trim_name	3.876415
trimId	3.860849
engine_type	3.352655
engine_cylinders	3.352655
fuel_type	2.757430
description	2.596665
transmission	2.139471
transmission_display	2.139471
exterior_color	1.665144
seller_rating	1.362382
body_type	0.451427
sp_id	0.003200
sp_name	0.000000
vin	0.000000
savings_amount	0.000000
price	0.000000
model_name	0.000000
make_name	0.000000
longitude	0.000000
listing_id	0.000000
listing_color	0.000000
listed_date	0.000000
latitude	0.000000
is_new	0.000000
franchise_dealer	0.000000
dealer_zip	0.000000
daysonmarket	0.000000
city	0.000000
year	0.000000

dtype: float64

1. feature selection

bed_height	85.696924
bed_length	85.696924
owner_count	50.566426
fleet	47.552533
theft_title	47.552533
isCab	47.552533
has_accidents	47.552533
frame_damaged	47.552533
salvage	47.552533
franchise_make	19.087579
torque	17.259537
highway_fuel_economy	16.375948

동일한 행 개수: 1426595

	vin	back_legroom	bed	bed_height	bed_length
0	ZACNJABB5KPJ92081	35.1 in	NaN	NaN	NaN
1	SALCJ2FX1LH858117	38.1 in	NaN	NaN	NaN
3	SALRR2RV0L2433391	37.6 in	NaN	NaN	NaN
4	SALCJ2FXXLH862327	38.1 in	NaN	NaN	NaN
6	3MZBPABL6KM107908	35.1 in	NaN	NaN	NaN
...	...	...	...	...	...
3000022	JN1BJ1CV4LW259601	33.4 in	NaN	NaN	NaN
3000024	1FMJK2AT4LEA69747	41.5 in	NaN	NaN	NaN
3000029	2C3CDZBT8LH222948	33.1 in	NaN	NaN	NaN
3000032	3N6CM0KNXLK704386	--	NaN	NaN	NaN
3000036	1GNERFKW0LJ225508	38.4 in	NaN	NaN	NaN

	body_type	cabin	city	city_fuel_economy
0	SUV / Crossover	NaN	Bayamon	NaN
1	SUV / Crossover	NaN	San Juan	NaN
3	SUV / Crossover	NaN	San Juan	NaN
4	SUV / Crossover	NaN	San Juan	NaN
6	Sedan	NaN	Bayamon	NaN
...	...	...	...	...
3000022	SUV / Crossover	NaN	Napa	25.0
3000024	SUV / Crossover	NaN	Ukiah	16.0
3000029	Coupe	NaN	Napa	16.0
3000032	Van	NaN	Napa	24.0
3000036	SUV / Crossover	NaN	Vallejo	18.0

1. feature selection

	combine_fuel_economy	...	transmission	\
0	NaN	...	A	
1	NaN	...	A	
3	NaN	...	A	
4	NaN	...	A	
6	NaN	...	A	
...	...	...	...	
3000022	NaN	...	CVT	
3000024	NaN	...	A	
3000029	NaN	...	A	
3000032	NaN	...	CVT	
3000036	NaN	...	A	

	transmission_display	trimId	trim_name	\
0	9-Speed Automatic Overdrive	t83804	Latitude FWD	
1	9-Speed Automatic Overdrive	t86759	S AWD	
3	8-Speed Automatic Overdrive	t86074	V6 HSE AWD	
4	9-Speed Automatic Overdrive	t86759	S AWD	
6	6-Speed Automatic Overdrive	t85256	Sedan FWD	
...	...	...	...	
3000022	Continuously Variable Transmission	t89964	SV FWD	
3000024	Automatic	t88161	Limited MAX 4WD	
3000029	8-Speed Automatic	t90116	R/T RWD	
3000032	Continuously Variable Transmission	t88980	S FWD	
3000036	Automatic	t85763	LS FWD	

	vehicle_damage_category	wheel_system	wheel_system_display	wheelbase	\
0	NaN	FWD	Front-Wheel Drive	101.2 in	
1	NaN	AWD	All-Wheel Drive	107.9 in	
3	NaN	AWD	All-Wheel Drive	115 in	
4	NaN	AWD	All-Wheel Drive	107.9 in	
6	NaN	FWD	Front-Wheel Drive	107.3 in	
...	...	...	...	...	
3000022	NaN	FWD	Front-Wheel Drive	104.2 in	
3000024	NaN	4WD	Four-Wheel Drive	131.6 in	
3000029	NaN	RWD	Rear-Wheel Drive	116 in	
3000032	NaN	FWD	Front-Wheel Drive	115.2 in	
3000036	NaN	FWD	Front-Wheel Drive	120.9 in	

	width	year
0	79.6 in	2019
1	85.6 in	2020
3	87.4 in	2020
4	85.6 in	2020
6	70.7 in	2019
...	...	...
3000022	72.3 in	2020
3000024	93.4 in	2020
3000029	85.4 in	2020
3000032	68.1 in	2020
3000036	78.6 in	2020

[1426595 rows x 66 columns]

1. feature selection

```python
# 동일한 결측치를 가진 행 삭제
df_cleaned = df[~missing_rows]

# 결과 확인
print(f"원본 데이터 크기: {df.shape}")
print(f"결측치 행 삭제 후 데이터 크기: {df_cleaned.shape}")
```

```
원본 데이터 크기: (3000040, 66)
결측치 행 삭제 후 데이터 크기: (1573445, 66)
```

+ Code + Markdown

1. feature selection

vin	ZACNJABB5KPJ92081 JF1VA2M67G9829723
dealer_zip	00969 00922
listing_id	265946296 173473508
sp_id	370467 389227

trimId	t86759 t86074
main_picture_url	https://static.cargurus.com/images/forsale/2020/05/15/18/25/2020_land_rover_discovery_sport-pic-3854...
sp_name	Land Rover San Juan Atlantic Chevrolet Cadillac
description	[!@@Additional Info@@!]Engine: 2.4L I4 ZERO EVAP M-AIR,Full Size Temporary Use Spare Tire,Manufactur...

1. feature selection-duplication

franchise_make vs make_name

franchise_make Type String. The company that owns the franchise.		make_name	
[null] Ford Other (2031884)	19% 13% 68%	Ford Chevrolet Other (2146812)	16% 13% 72%
Jeep		Jeep	
Land Rover		Land Rover	
FIAT		Subaru	
Land Rover		Land Rover	
Land Rover		Land Rover	

- 겹치지만 franchise_make가 결측치가 많아 make_name을 select

wheel_system vs wheel_System_display

wheel_system		wheel_system_display	
FWD AWD Other (1042942)	42% 23% 35%	Front-Wheel Drive All-Wheel Drive Other (1042942)	42% 23% 35%
FWD		Front-Wheel Drive	
AWD		All-Wheel Drive	
AWD		All-Wheel Drive	
AWD		All-Wheel Drive	
AWD		All-Wheel Drive	
AWD		All-Wheel Drive	
FWD		Front-Wheel Drive	
AWD		All-Wheel Drive	
AWD		All-Wheel Drive	
AWD		All-Wheel Drive	
RWD		Rear-Wheel Drive	

- wheel_system이 코드화 된 내용이기에 select

1. feature selection-duplication

trim_name

trim_name	
[null]	4%
SE FWD	3%
Other (2796126)	93%
Latitude FWD	
S AWD	
Base	
V6 HSE AWD	
S AWD	

- trim_name이 다른 feature값을 포함

exterior_color vs listing_color

exterior_color Type String. Exterior dominant color of the vehicle		listing_color Dominant color group from the exterior color.	
Black	5%	WHITE	22%
White	4%	BLACK	20%
Other (2711017)	90%	Other (1745477)	58%
Solar Yellow		YELLOW	
Narvik Black		BLACK	
None		UNKNOWN	
Eiger Gray		GRAY	
Narvik Black		BLACK	
Kaikoura Stone		UNKNOWN	
SONIC SILVER		SILVER	
Fuji White		WHITE	
Eiger Gray		GRAY	

- exterior_color을 단순화 시킨 것이 listing_color

1. feature selection-duplication

savings_amount

savings_amount

0
0
0
1749
0
1861

- savings_amount는 판매자가 수정하여 내린 가격편차임. 중요한 것은 거래된 가격임.

transmission vs transmission_display

transmission		transmission_display	
A	81%	Automatic	42%
CVT	15%	Continuously Vari...	15%
Other (123679)	4%	Other (1289975)	43%
A		9-Speed Automatic Overdrive	
A		9-Speed Automatic Overdrive	
M		6-Speed Manual	

- transmission_display가 transmission을 포함

1. feature selection-TAs

franchise_dealer

✓ franchise_dealer

Type Boolean. Whether the dealer is a franchise dealer.

true
2.43m 81%

false
569k 19%

True
True
True
True
True

- 인증된 딜러인지 아닌지

is_new

✓ is_new

If True means the vehicle was launched less than 2 years before 2020.

true
1.47m 49%

false
1.53m 51%

True
True
False
True
True
True

- 이 차가 출고 후 2년 이하

maximum_seating

A maximum_seating

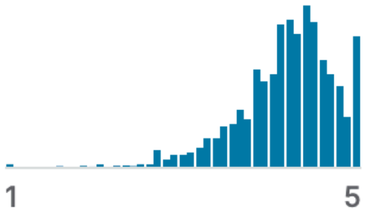
5 seats	63%
7 seats	10%
Other (819353)	27%

5 seats
7 seats
5 seats
7 seats
7 seats
5 seats
5 seats
5 seats

- 최대 좌석수. 세단은 5석, suv는 7석 등 차량 종류에 영향

1. feature selection-Survey

seller_rating

seller_rating

2.8
3.0
3.0
3.0
3.0
2.8
3.0
3.0

- 평점은 주관적인 데이터라서 다루기가 어려움

1. feature selection

sold_date = listed_date + daysonmarket 이라고 할 때
sold_date < 2020-09-13 이면 차량이 팔렸다고 간주

listed_date	daysonmarket
The date the vehicle was listed on the website.	Type Integer. Days since the vehicle was first listed on the website.
 2010-11-03 2020-09-13	 0 3599
2019-04-06	522
2020-02-15	207

```python
# 날짜 형식 변환
df['listed_date'] = pd.to_datetime(df['listed_date'])

# 수집 종료일 정의
end_date = pd.to_datetime('2020-09-13')

# 차량 판매일 계산
df['sold_date'] = df['listed_date'] + pd.to_timedelta(df['daysonmarket'], unit='D')

# 팔린 차량 추출
sold_cars = df[df['sold_date'] <= end_date]

# 결과 확인
print(f"총 팔린 차량 수: {len(sold_cars)}")
print(sold_cars.head())
```

```
총 팔린 차량 수: 2986075
                vin back_legroom  bed bed_height bed_length        body_type  \
0  ZACNJABB5KPJ92081      35.1 in  NaN        NaN        NaN  SUV / Crossover
1  SALCJ2FX1LH858117      38.1 in  NaN        NaN        NaN  SUV / Crossover
2  JF1VA2M67G9829723      35.4 in  NaN        NaN        NaN            Sedan
3  SALRR2RV0L2433391      37.6 in  NaN        NaN        NaN  SUV / Crossover
4  SALCJ2FXXLH862327      38.1 in  NaN        NaN        NaN  SUV / Crossover

  cabin      city  city_fuel_economy  combine_fuel_economy  ...  \
0   NaN   Bayamon                NaN                   NaN  ...
1   NaN  San Juan                NaN                   NaN  ...
2   NaN  Guaynabo               17.0                   NaN  ...
3   NaN  San Juan                NaN                   NaN  ...
4   NaN  San Juan                NaN                   NaN  ...
```

1. feature selection

torque	
[null]	17%
383 lb-ft @ 4,100 ...	3%
Other (2393947)	80%
200 lb-ft @ 1,750 RPM	
269 lb-ft @ 1,400 RPM	

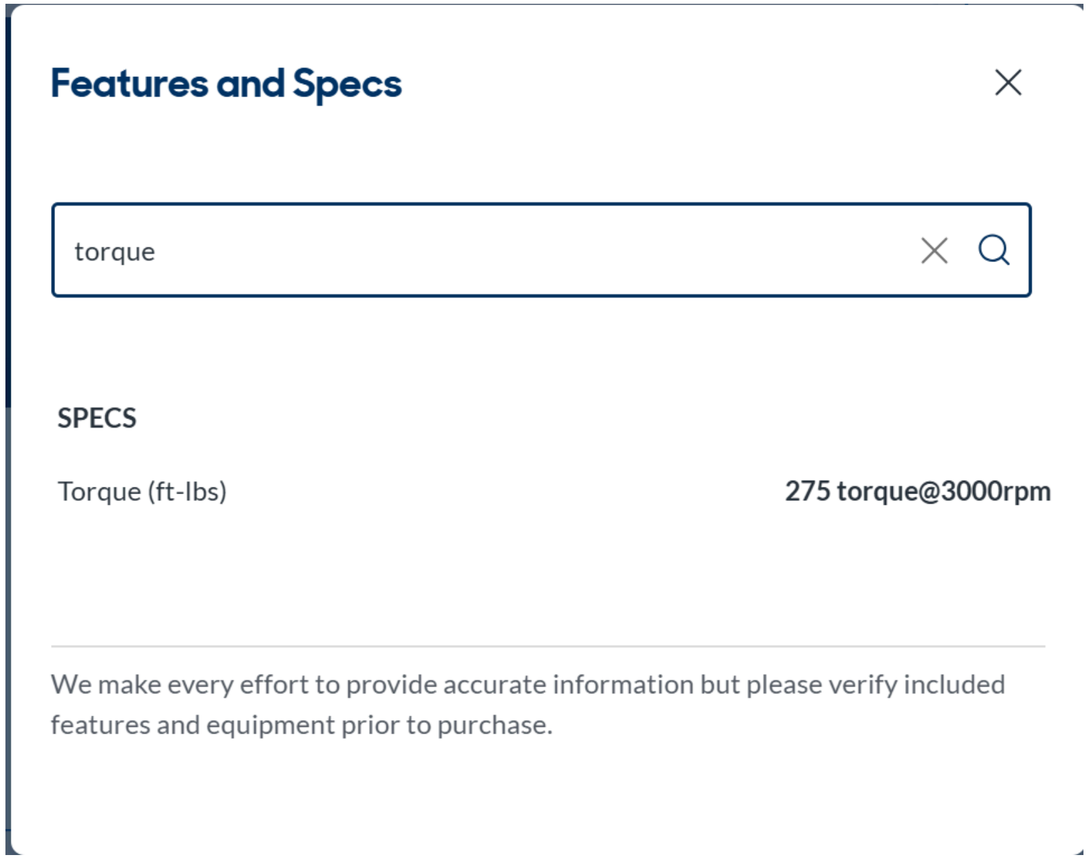
power	
[null]	16%
355 hp @ 5,600 R...	3%
Other (2430314)	81%
177 hp @ 5,750 RPM	
246 hp @ 5,500 RPM	

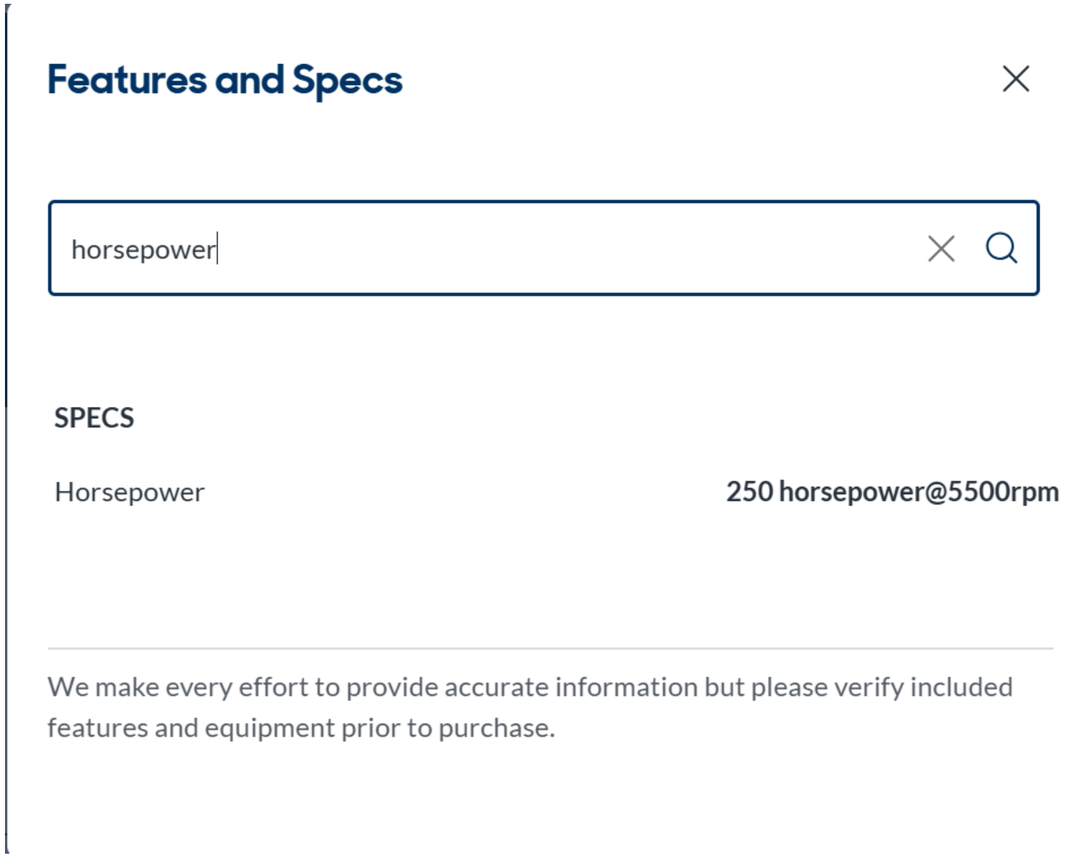
torque vs power

```
print(df[['torque_rpm','power_rpm']])
```

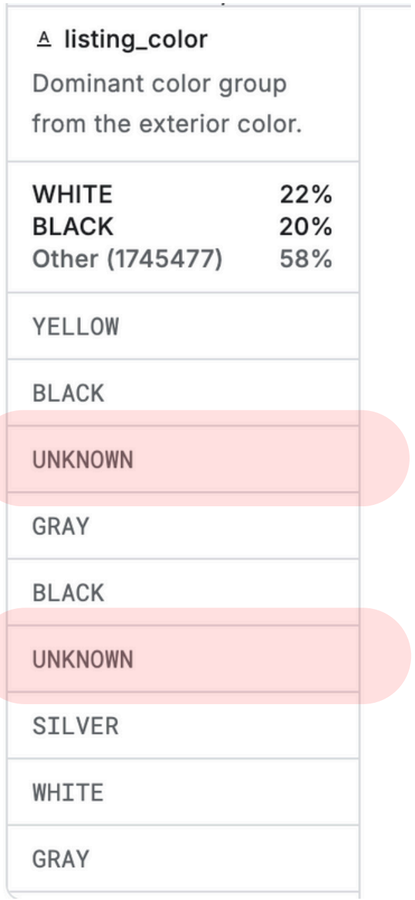
	torque_rpm	power_rpm
0	1750.0	5750.0
1	1400.0	5500.0
2	4000.0	6000.0
3	3500.0	6500.0
4	1400.0	5500.0
...	...	...
3000035	NaN	NaN
3000036	2800.0	6800.0
3000037	1750.0	5500.0
3000038	1750.0	4000.0
3000039	4400.0	6000.0

1. feature selection





2. Missing value



< 결측치 처리 방법 >

범주형 데이터 : 결측치를 'unknown'으로 대체(하나의 새로운 범주로 취급)

수치형 데이터 : 결측치를 중앙값으로 대체

boolean : 결측치를 'false'로 대체

3. 1st preprocessing

1/28 PREPROCESSING

3. 1st preprocessing

bed_height	85.696924
bed_length	85.696924
owner_count	50.566426
fleet	47.552533
theft_title	47.552533
isCab	47.552533
has_accidents	47.552533
frame_damaged	47.552533
salvage	47.552533
franchise_make	19.087579
torque	17.259537
highway_fuel_economy	16.375948

결측치 data sample
삭제 후 >>

각 feature의 결측치 비율 (%):

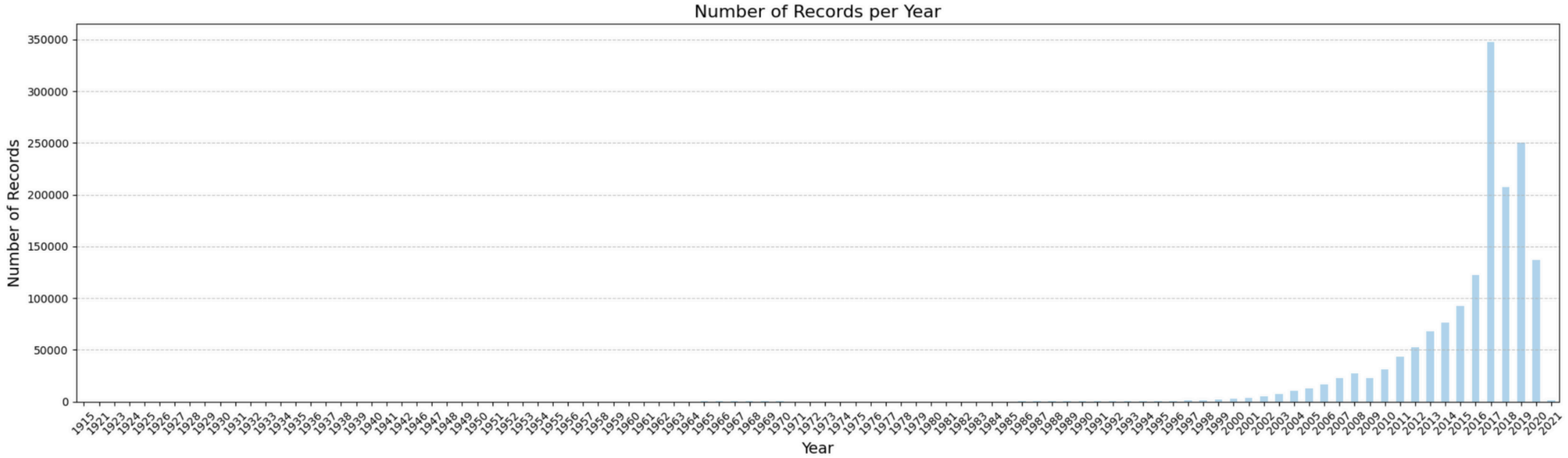
highway_fuel_economy	15.697848
city_fuel_economy	15.697848
interior_color	15.496697
torque	12.825742
owner_count	5.746499
major_options	4.510549
maximum_seating	4.238470
engine_displacement	4.062741
horsepower	4.062741
wheel_system	3.237609
trim_name	2.830477
engine_type	2.542320
exterior_color	2.244947
fuel_type	2.070171
transmission	1.702697
mileage	1.232328
body_type	0.118339
savings_amount	0.000000
salvage	0.000000
price	0.000000
theft_title	0.000000
model_name	0.000000
listed_date	0.000000
make_name	0.000000
listing_color	0.000000
city	0.000000
isCab	0.000000
has_accidents	0.000000
franchise_dealer	0.000000
frame_damaged	0.000000
fleet	0.000000
dealer_zip	0.000000
daysonmarket	0.000000
year	0.000000

dtype: float64

3. 1st preprocessing

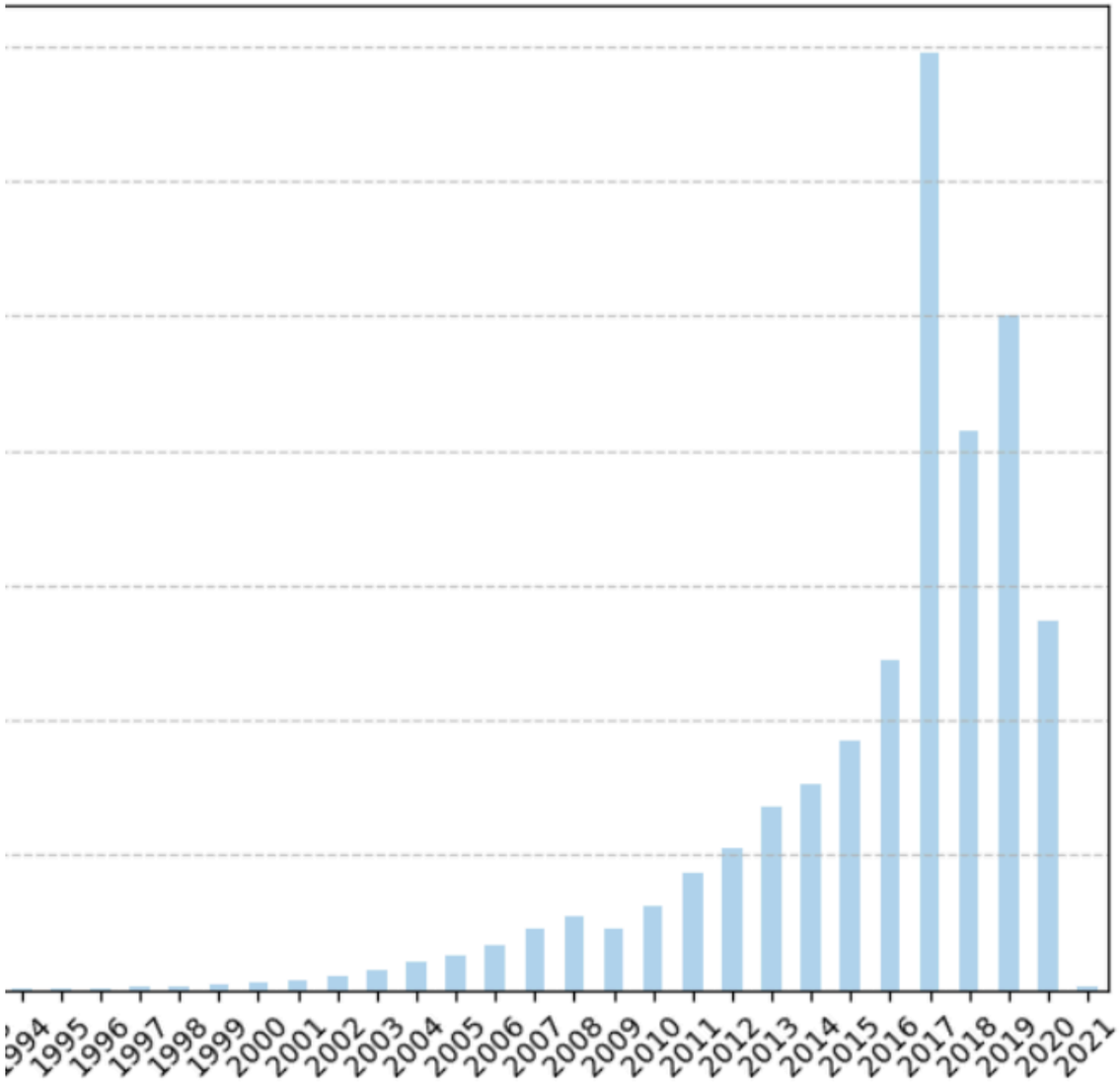
```
<class 'pandas.core.frame.DataFrame'>
Index: 1573445 entries, 2 to 3000039
Data columns (total 34 columns):
 #   Column                Non-Null Count    Dtype
---  ------                --------------    -----
 0   body_type             1571583 non-null  object
 1   city                  1573445 non-null  object
 2   city_fuel_economy     1326448 non-null  Int64
 3   daysonmarket          1573445 non-null  int64
 4   dealer_zip            1573445 non-null  object
 5   engine_displacement   1509520 non-null  Int64
 6   engine_type           1533443 non-null  object
 7   exterior_color        1538122 non-null  object
 8   fleet                 1573445 non-null  object
 9   frame_damaged         1573445 non-null  object
 10  franchise_dealer      1573445 non-null  bool
 11  fuel_type             1540872 non-null  object
 12  has_accidents         1573445 non-null  object
 13  highway_fuel_economy  1326448 non-null  Int64
 14  horsepower            1509520 non-null  Int64
 15  interior_color        1329613 non-null  object
 16  isCab                 1573445 non-null  object
 17  listed_date           1573445 non-null  datetime64[ns]
 18  listing_color         1573445 non-null  object
 19  major_options         1502474 non-null  object
 20  make_name             1573445 non-null  object
 21  maximum_seating       1506281 non-null  object
 22  mileage               1554055 non-null  Int64
 23  model_name            1573445 non-null  object
 24  owner_count           1483027 non-null  Int64
 25  price                 1573445 non-null  int64
 26  salvage               1573445 non-null  object
 27  savings_amount        1573445 non-null  int64
 28  theft_title           1573445 non-null  object
 29  torque                1371639 non-null  object
 30  transmission          1546654 non-null  object
 31  trim_name             1528909 non-null  object
 32  wheel_system          1522503 non-null  object
 33  year                  1573445 non-null  int64
dtypes: Int64(6), bool(1), datetime64[ns](1), int64(4), object(22)
memory usage: 418.7+ MB
```

3. 1st preprocessing



year별 차량 data sample counts graph

3. 1st preprocessing

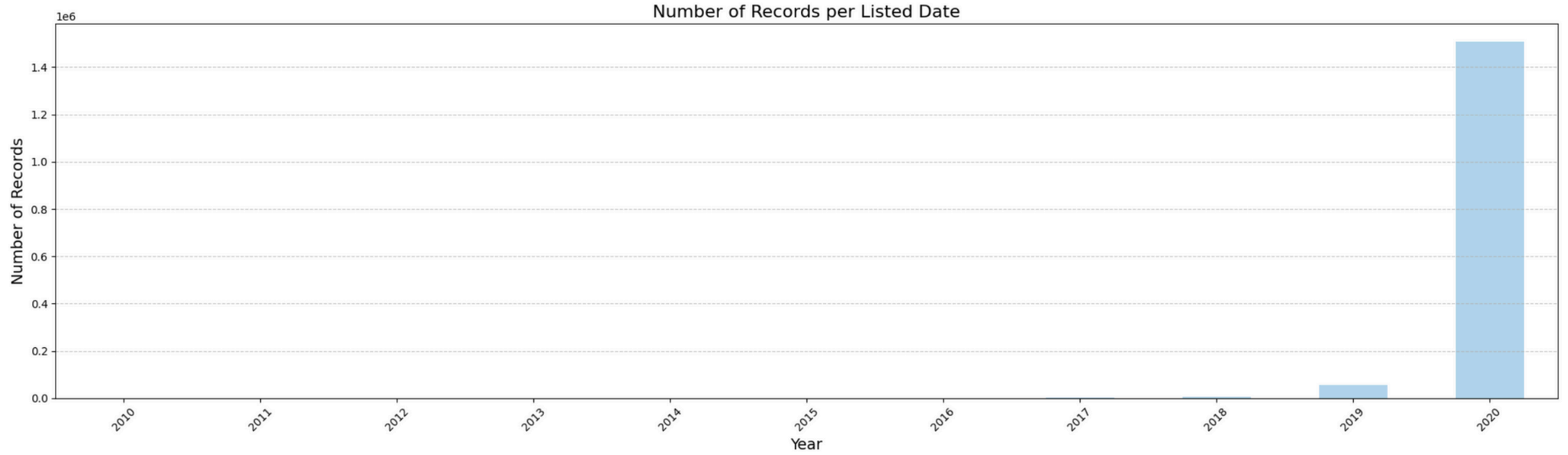


3. 1st preprocessing

listed date별 차량 data sample counts graph

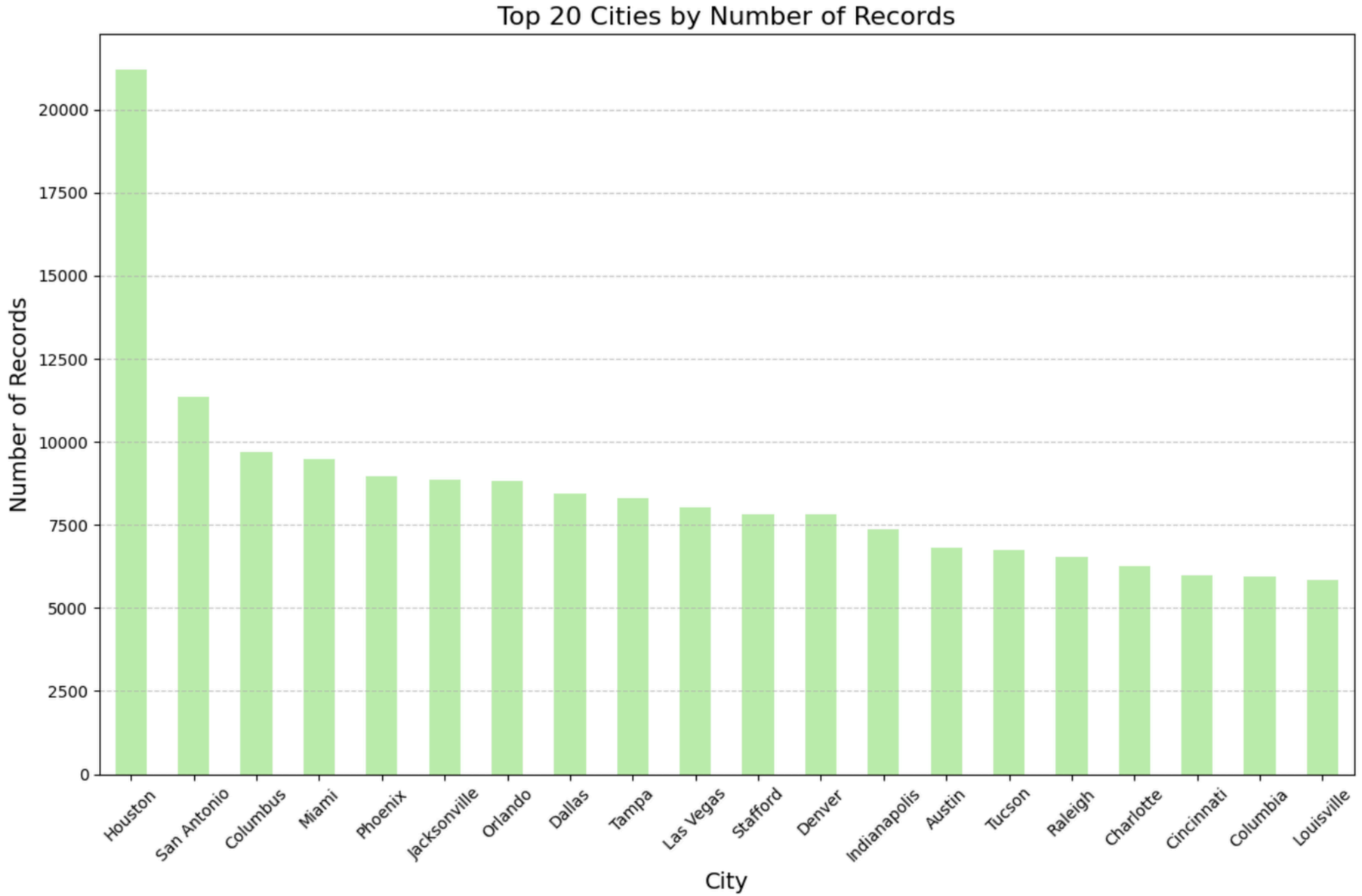
[16]:

```python
import matplotlib.pyplot as plt

# 'listed_date'에서 연도만 추출하여 새로운 컬럼 'year' 생성
df['year'] = df['listed_date'].astype(str).str[:4]

# 연도별 데이터 개수 계산
year_counts = df['year'].value_counts().sort_index()

# 연도별 데이터 개수 시각화
plt.figure(figsize=(20, 6))
year_counts.plot(kind='bar', color='skyblue', alpha=0.8)
plt.title('Number of Records per Listed Date', fontsize=16)
plt.xlabel('Year', fontsize=14)
plt.ylabel('Number of Records', fontsize=14)
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

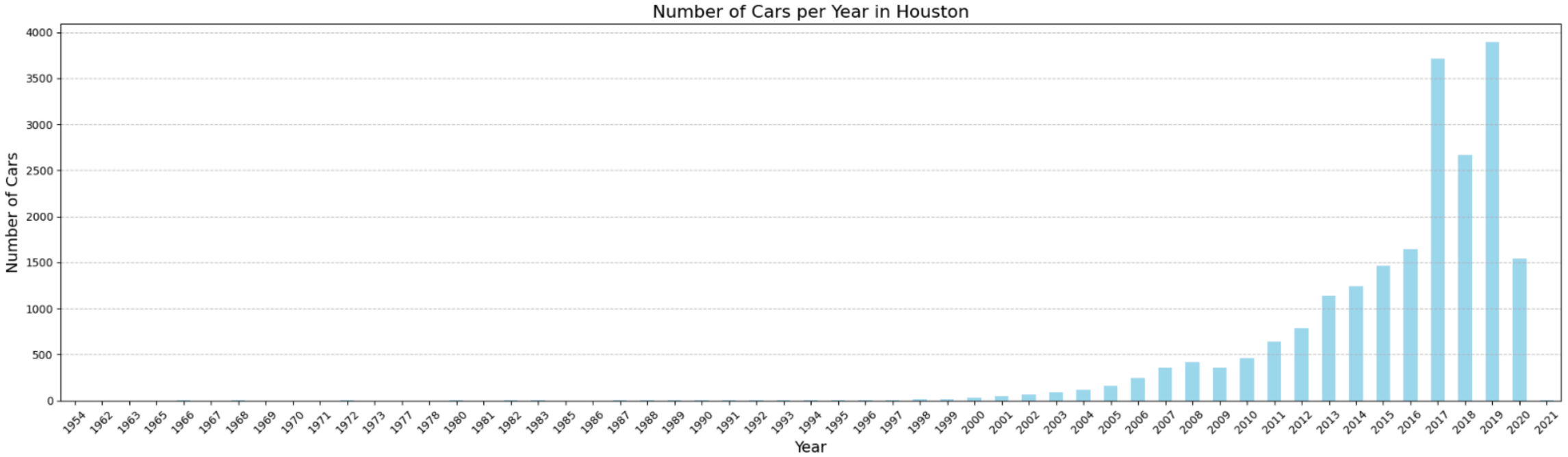


3. 1st preprocessing

city별 차량 data sample counts graph



3. 1st preprocessing



3. 1st preprocessing

```python
from collections import Counter
import pandas as pd
import ast  # 안전한 문자열 → 리스트 변환

# 데이터 샘플링: 전체의 10%만 사용하여 메모리 절약
df_sample = df.sample(frac=0.1, random_state=42)  # 10% 샘플 사용

# major_options 컬럼에서 리스트를 풀어서 한 개의 리스트로 만듦
all_options = []
for row in df_sample['major_options'].dropna():
    try:
        options = ast.literal_eval(row)  # 문자열을 안전하게 리스트로 변환
        if isinstance(options, list):
            all_options.extend(options)  # 리스트 내부 요소 개별 추가
    except:
        continue  # 변환 실패 시 무시 (예외 방지)

# 각 옵션별 빈도수 계산
option_counts = Counter(all_options)

# 결과를 DataFrame으로 변환 후 정렬하여 출력
option_counts_df = pd.DataFrame(option_counts.items(), columns=['Option', 'Count']).sort_values(by='Count', ascending=False)

# 모든 행 출력 설정 (기본값은 10~20개만 출력)
pd.set_option('display.max_rows', None)

display(option_counts_df)
```

3. 1st preprocessing

major options value counts

	Option	Count
11	Bluetooth	114149
12	Backup Camera	113149
5	Alloy Wheels	107720
10	Heated Seats	66047
3	Navigation System	60560
1	Sunroof/Moonroof	50015
0	Leather Seats	49957
18	Remote Start	45689
7	Blind Spot Monitoring	36890
14	CarPlay	28110
8	Parking Sensors	24677
44	Android Auto	23317
6	Third Row Seating	18526
35	Steel Wheels	16653
4	Adaptive Cruise Control	16598
9	Premium Package	9890
19	Quick Order Package	9888
2	Power Package	7306
29	Tow Package	6803
15	Convenience Package	6503
38	Multi Zone Climate Control	5642
22	Premium Wheels	5605
21	Technology Package	4939
23	Heat Package	4497
51	Chrome Wheels	4091
24	Cold Weather Package	4001
43	Appearance Package	3962
17	Preferred Package	3962
40	Suspension Package	3622
20	Sport Package	3363
13	SE Package	2939
37	Adaptive Suspension	2566
25	LE Package	2521
47	Trailer Package	2264
33	Memory Package	2227
34	Comfort Package	2043
54	Luxury Package	1911
55	Off Road Package	1903
36	Driver Assistance Package	1566
28	DVD Entertainment System	1488
26	Cargo Package	1333
66	Audio Package	1324
46	LT Package	1214
41	Premium Plus Package	1152
67	Sound Package	1126
68	Driver Confidence Package	1063
48	302A Luxury Equipment Group	952
57	Value Package	943
45	XLT Package	878
56	Premium Smooth Ride Suspension Package	875
50	Premium 1 Package	828
61	RS Package	703
27	M Sport Package	609
73	Safety Package	553
58	SL Package	534
60	301A Mid Equipment Group	521
32	Trim Package	511
62	Light Package	488
53	Security Package	476

3. 1st preprocessing

```
unique_filtered_packages
```

	make_name	model_name	filtered_packages
6	Hyundai	Elantra	Sport Package
6	Hyundai	Elantra	Audio Package
6	Hyundai	Elantra	Heat Package
6	Hyundai	Elantra	Premium Package
7	Chevrolet	Malibu	Driver Confidence Package
...	...	...	...
1573288	Volkswagen	Jetta Hybrid	Premium Audio Package
1573288	Volkswagen	Jetta Hybrid	Heat Package
1573288	Volkswagen	Jetta Hybrid	Audio Package
1573352	GMC	Yukon Hybrid	Comfort Package
1573352	GMC	Yukon Hybrid	Convenience Package

11732 rows × 3 columns

3. 1st preprocessing

engine_type vs fuel_type

fuel_type 값 분포:

fuel_type	
Gasoline	2598436
Flex Fuel Vehicle	155993
Hybrid	76012
Diesel	44452
Biodiesel	25855
Electric	16416
Compressed Natural Gas	146
Propane	6

engine_type 값 분포:

engine_type	
I4	1418291
V6	737066
V8	279939
V8 Flex Fuel Vehicle	78372
I4 Hybrid	72712
V6 Flex Fuel Vehicle	68337
H4	65861
I3	54301
I6	28322
I6 Diesel	23438
V8 Biodiesel	22148
I4 Flex Fuel Vehicle	9187
I4 Diesel	7210
V6 Diesel	6657
V8 Diesel	6266
I5	5301
H6	4686
V6 Biodiesel	3611
V6 Hybrid	3058
V12	1316
V10	1291
I2	897
W12	484
V8 Hybrid	127
W12 Flex Fuel Vehicle	97
I5 Biodiesel	96
V8 Compressed Natural Gas	95
H4 Hybrid	94
R2	65
I5 Diesel	49
I4 Compressed Natural Gas	48
I6 Hybrid	16
V8 Propane	6
W8	3
I3 Hybrid	3
V6 Compressed Natural Gas	3
V10 Diesel	2
W16	2
V12 Hybrid	2

모든 row에서 engine type과 fuel type 뒤의 값이 같은가?

3. 1st preprocessing

```python
# engine_fuel_match 결과 요약
total_rows = len(df)
true_count = df['engine_fuel_match'].sum()  # True의 개수
false_count = total_rows - true_count  # False의 개수

print(f"Total Rows: {total_rows}")
print(f"True Count: {true_count} ({true_count / total_rows * 100:.2f}%)")
print(f"False Count: {false_count} ({false_count / total_rows * 100:.2f}%)")
```

```
Total Rows: 3000040
True Count: 3000040 (100.00%)
False Count: 0 (0.00%)
```

engine_type vs fuel_type

3. 1st preprocessing

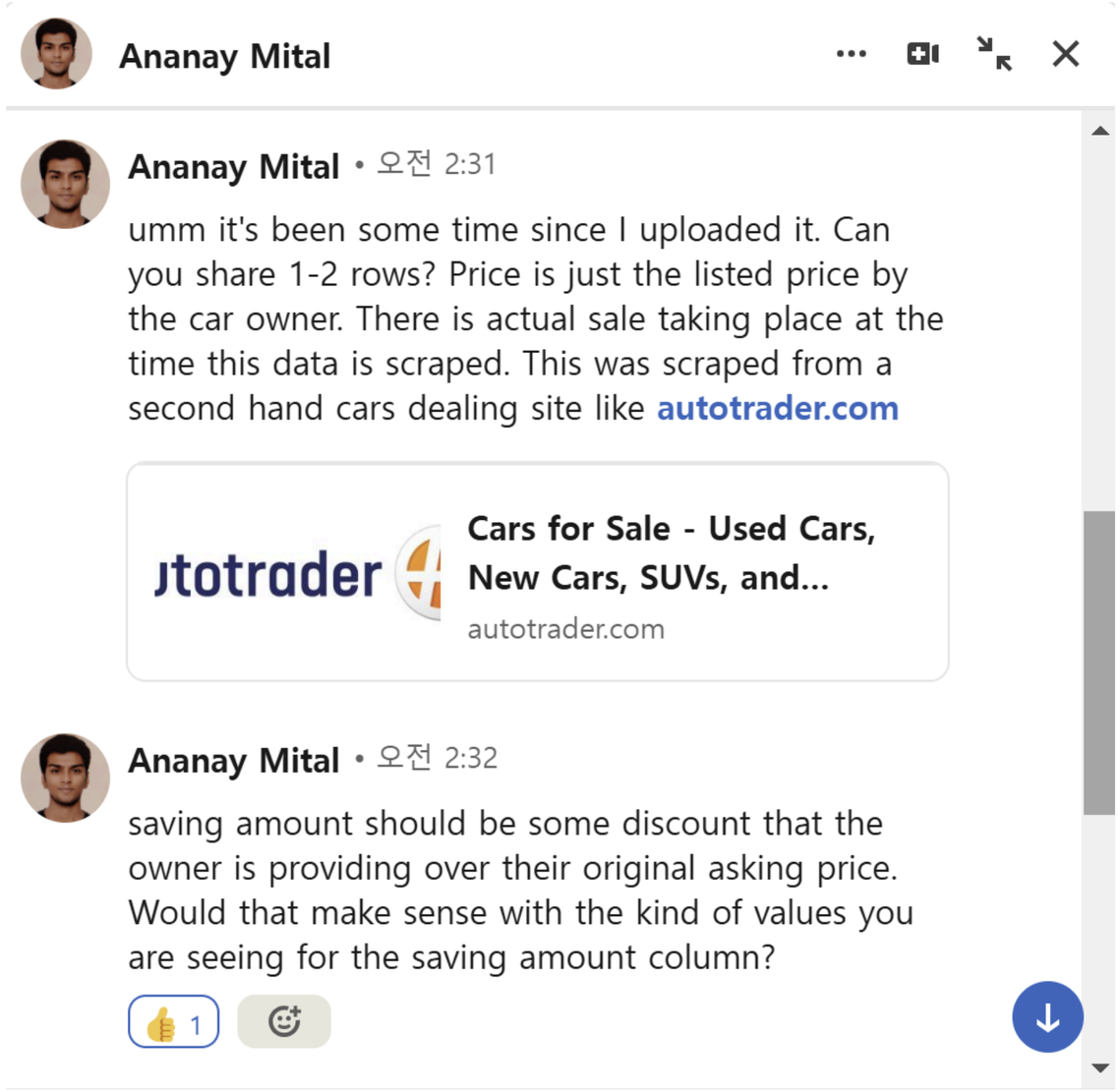
Interior_color vs listing_color

```
interior_color 값 분포:
interior_color
Black                                                    871393
Gray                                                     195908
Jet Black                                                186195
Black (Ebony)                                            142839
Black (Charcoal)                                         112051
                                                          ...
Brown (Cappuccino w/Heated Lincoln Soft Touch Front Seats)     1
Circuit Red Nuluxe[nuluxe] With Dark Gray Streamli             1
Nut Brown/ Black Leather                                       1
Black/Orange w/Fabric Seat Trim (FD)                           1
Brown (Espresso/Iv/Tan/Esp/Iv/Iv)                              1
```

```
Name: count, Length: 1749,
listing_color 값 분포:
listing_color
WHITE      666564
BLACK      587999
UNKNOWN    399905
SILVER     384779
GRAY       377442
RED        252917
BLUE       249758
GREEN       24074
BROWN       22611
ORANGE      11631
GOLD        10297
TEAL         5453
YELLOW       5003
PURPLE       1468
PINK          139
```

3. 1st preprocessing

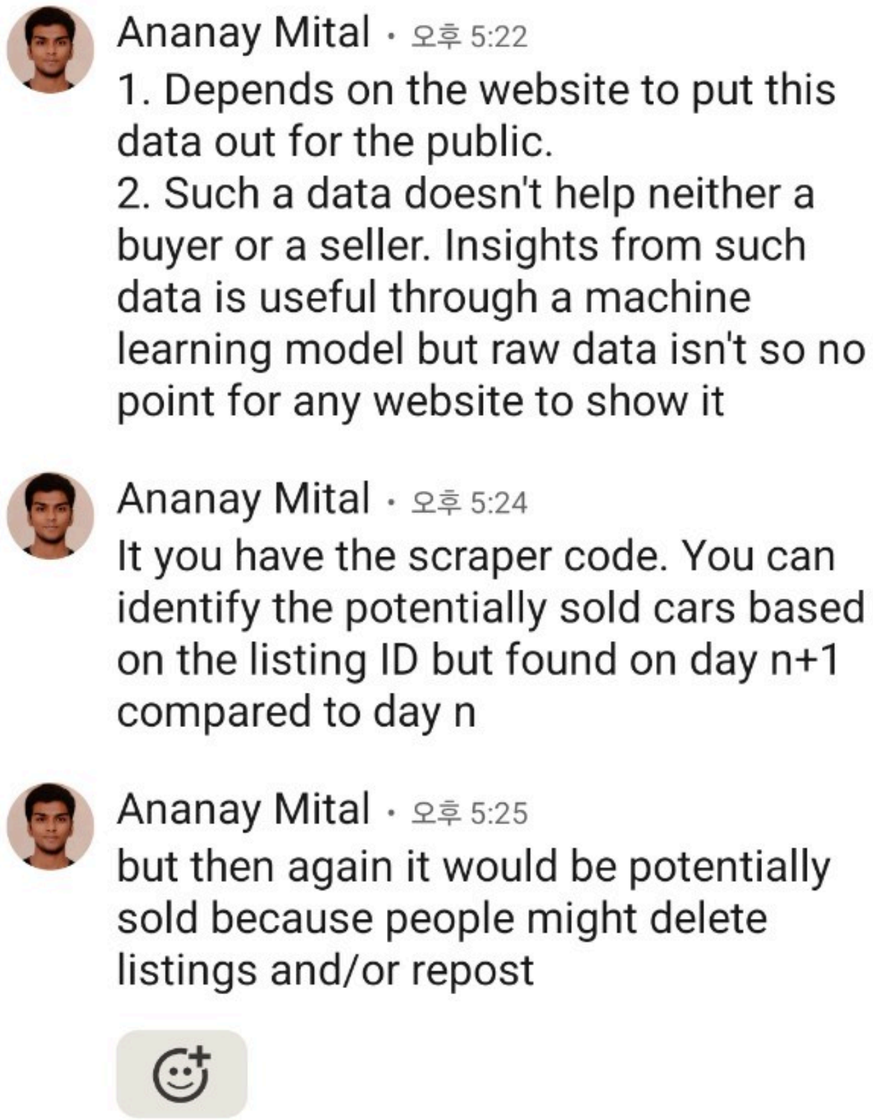
price & saving amount feature



price : **중고차 사이트에 올린 가격**

saving amount : **처음에 올린 기존 가격 보다 감소된 가격(가격을 낮췄을 경우)**

3. 1st preprocessing



whether the car is sold or not

- 원본 데이터 자체는 의미가 없고, 머신러닝 모델을 통해 분석하여 유용한 인사이트를 얻을 수 있음
- 스크래핑 코드를 사용하면 list ID를 기반으로 특정 차량이 특정 날짜 (n일)에 존재했는지 확인하고, 그 다음날 (n+1일)에 해당 ID가 사라졌다면 해당 차량이 판매되었을 가능성
- 그러나 사람들이 단순히 차량 목록을 삭제하거나 다시 올릴 수도 있기에, 단순히 listID가 사라졌다고 해서 반드시 판매되었다고 확신은 불가

3. 1st preprocessing

1/29 PREPROCESSING

3. 1st preprocessing

Interior_color vs listing_color

```
interior_color 값 분포:
interior_color
Black                                                    871393
Gray                                                     195908
Jet Black                                                186195
Black (Ebony)                                            142839
Black (Charcoal)                                         112051
                                                          ...
Brown (Cappuccino w/Heated Lincoln Soft Touch Front Seats)     1
Circuit Red Nuluxe[nuluxe] With Dark Gray Streamli             1
Nut Brown/ Black Leather                                       1
Black/Orange w/Fabric Seat Trim (FD)                           1
Brown (Espresso/Iv/Tan/Esp/Iv/Iv)                              1
```

```
Name: count, Length: 1749,
listing_color 값 분포:
listing_color
WHITE      666564
BLACK      587999
UNKNOWN    399905
SILVER     384779
GRAY       377442
RED        252917
BLUE       249758
GREEN       24074
BROWN       22611
ORANGE      11631
GOLD        10297
TEAL         5453
YELLOW       5003
PURPLE       1468
PINK          139
```