Car Analysis

March 24, 2024

1 Car prices Dataset

- Link to dataset - https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data

[1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

[2]:
```
df = pd.read_csv('car_prices.csv')
```

[3]:
```
df.head()
```

[3]:

	year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior	seller	mmr	sellingprice	saledate
0	2015	Kia	Sorento	LX	SUV	automatic	5xyktca69fg566472	ca	5.0	16639.0	white	black	kia motors america inc	20500.0	21500.0	Tue Dec 16 2014 12:30:00 GMT-0800 (PST)
1	2015	Kia	Sorento	LX	SUV	automatic	5xyktca69fg561319	ca	5.0	9393.0	white	beige	kia motors america inc	20800.0	21500.0	Tue Dec 16 2014 12:30:00 GMT-0800 (PST)
2	2014	BMW	3 Series	328i SULEV	Sedan	automatic	wba3c1c51ek116351	ca	45.0	1331.0	gray	black	financial services remarketing (lease)	31900.0	30000.0	Thu Jan 15 2015 04:30:00 GMT-0800 (PST)
3	2015	Volvo	S60	T5	Sedan	automatic	yv1612tb4f1310987	ca	41.0	14282.0	white	black	volvo na rep/world omni	27500.0	27750.0	Thu Jan 29 2015 04:30:00 GMT-0800 (PST)
4	2014	BMW	6 Series Gran Coupe	650i	Sedan	automatic	wba6b2c57ed129731	ca	43.0	2641.0	gray	black	financial services remarketing (lease)	66000.0	67000.0	

```
4  Thu Dec 18 2014 12:30:00 GMT-0800 (PST)
```

[4]:
```
df.describe()
```

[4]:
```
                year      condition       odometer            mmr  \
count  558837.000000  547017.000000  558743.000000  558799.000000
mean     2010.038927      30.672365   68320.017767   13769.377495
std         3.966864      13.402832   53398.542821    9679.967174
min      1982.000000       1.000000       1.000000      25.000000
25%      2007.000000      23.000000   28371.000000    7100.000000
50%      2012.000000      35.000000   52254.000000   12250.000000
75%      2013.000000      42.000000   99109.000000   18300.000000
max      2015.000000      49.000000  999999.000000  182000.000000

        sellingprice
count  558825.000000
mean    13611.358810
std      9749.501628
min         1.000000
25%      6900.000000
50%     12100.000000
75%     18200.000000
max    230000.000000
```

[5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 558837 entries, 0 to 558836
Data columns (total 16 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   year          558837 non-null  int64
 1   make          548536 non-null  object
 2   model         548438 non-null  object
 3   trim          548186 non-null  object
 4   body          545642 non-null  object
 5   transmission  493485 non-null  object
 6   vin           558833 non-null  object
 7   state         558837 non-null  object
 8   condition     547017 non-null  float64
 9   odometer      558743 non-null  float64
 10  color         558088 non-null  object
 11  interior      558088 non-null  object
 12  seller        558837 non-null  object
 13  mmr           558799 non-null  float64
 14  sellingprice  558825 non-null  float64
 15  saledate      558825 non-null  object
dtypes: float64(4), int64(1), object(11)
```

```
memory usage: 68.2+ MB
```

[6]:
```
df.shape[0]
```

[6]:
```
558837
```

1.1 Columns examination

[7]:
```
df.isna().sum()
```

[7]:
```
year              0
make          10301
model         10399
trim          10651
body          13195
transmission  65352
vin               4
state             0
condition     11820
odometer         94
color           749
interior        749
seller            0
mmr              38
sellingprice     12
saledate         12
dtype: int64
```

[8]:
```
df.dropna(subset = ["make","model","trim","color","odometer"],inplace = True)
```

[9]:
```
df.drop(columns = 'vin',inplace = True)
```

[10]:
```
df.isna().sum()
```

[10]:
```
year              0
make              0
model             0
trim              0
body           2535
transmission  63379
state             0
condition     11510
odometer          0
color             0
interior          0
seller            0
mmr              38
sellingprice     12
```

```
saledate           12
dtype: int64
```

2 Adjusting transmission column

[11]:
```
df["transmission"].value_counts()
```

[11]:
```
transmission
automatic    466955
manual        16918
sedan            15
Sedan            11
Name: count, dtype: int64
```

[12]:
```
df = df[~df["transmission"].isin(["sedan","Sedan"])]
df['transmission'] = df["transmission"].str.capitalize()
```

[13]:
```
df[(~df["transmission"].isna())].sort_values(by = ['make'])
```

[13]:
```
        year   make   model      trim   body transmission state  condition  \
271080  2001  Acura     MDX   Touring    SUV    Automatic    va       19.0
257571  2011  Acura      TL      Base  Sedan    Automatic    fl        2.0
443519  2007  Acura     RDX      Base    SUV    Automatic    az       32.0
148562  2011  Acura     MDX      Base    SUV    Automatic    il        3.0
16899   2007  Acura     MDX      Base    SUV    Automatic    tx        NaN
…       …     …         …        …       …      …            …
69157   2014     vw  routan   sel prm    NaN    Automatic    oh       43.0
69154   2014     vw  routan        se    NaN    Automatic    oh        5.0
103699  2013     vw  routan  se w/nav    NaN    Automatic    il        5.0
107814  2014     vw  routan        se    NaN    Automatic    il       39.0
103704  2013     vw  routan  se w/rse    NaN    Automatic    il       38.0

        odometer   color interior  \
271080  192889.0  silver    black
257571   72452.0    gray    beige
443519   95782.0   black     gray
148562   44254.0   black     gray
16899    99533.0   green         -
…          …      …        …
69157    25053.0       -     gray
69154     6842.0       -     gray
103699   10313.0       -    black
107814   15376.0       -     gray
103704   17502.0       -     gray

                                                seller      mmr  sellingprice  \
```

```
271080        credit acceptance corp/vrs/southfield   2975.0        2400.0
257571                     capital one auto finance  16000.0       16600.0
443519                       moore automotive group  10700.0        9800.0
148562  ahfc/honda lease/trust/hvt  inc. eot acuras  23500.0       22000.0
16899                             bmw of san antonio  13200.0       13200.0
…                                                  …       …           …
69157                                      vw credit  22300.0       25000.0
69154                                      vw credit  21900.0       21700.0
103699                                     vw credit  20800.0       20600.0
107814                                     vw credit  20800.0       20400.0
103704                                     vw credit  20200.0       19200.0

                                      saledate
271080  Thu Feb 12 2015 01:05:00 GMT-0800 (PST)
257571  Wed Feb 11 2015 01:10:00 GMT-0800 (PST)
443519  Thu May 21 2015 05:00:00 GMT-0700 (PDT)
148562  Thu Jan 22 2015 02:00:00 GMT-0800 (PST)
16899   Fri Dec 19 2014 10:00:00 GMT-0800 (PST)
…                                          …
69157   Tue Jan 06 2015 01:30:00 GMT-0800 (PST)
69154   Tue Jan 06 2015 01:30:00 GMT-0800 (PST)
103699  Tue Jan 13 2015 10:00:00 GMT-0800 (PST)
107814  Tue Jan 13 2015 10:00:00 GMT-0800 (PST)
103704  Tue Jan 13 2015 10:00:00 GMT-0800 (PST)

[483873 rows x 15 columns]
```

[14]:
```
df[df["transmission"].isna()].sort_values(by = ['make'])
```

[14]:
```
        year   make  model        trim       body transmission state  \
315386  2012  Acura    RDX        Base        SUV          NaN    pa
264896  2007  Acura    MDX        Base        SUV          NaN    pa
427777  2006  Acura    RSX        Base  Hatchback          NaN    pa
427829  2005  Acura     TL         3.2      Sedan          NaN    tx
147329  2012  Acura    MDX        Base        SUV          NaN    pa
…       …     …    …          …         …           …   …
65502   2011     vw  jetta  comfrtline        NaN          NaN    on
65518   2011     vw  jetta  comfrtline        NaN          NaN    on
65522   2011     vw  jetta  comfrtline        NaN          NaN    on
65526   2011     vw  jetta  comfrtline        NaN          NaN    on
27667   2002     vw  jetta    gls 1.8t        NaN          NaN    on

        condition  odometer   color interior  \
315386       48.0   32598.0   white     gray
264896       31.0  396211.0   black    beige
427777       19.0  117683.0   black    black
427829       22.0  126683.0   green    brown
```

```
147329       25.0   38162.0   black    black
…             …       …      …        …
65502        19.0   63397.0   brown    beige
65518         2.0   74093.0   white    black
65522        32.0  110754.0    blue    black
65526        44.0   60834.0  silver    black
27667         NaN  123304.0   black        -

                                              seller      mmr  sellingprice  \
315386                          adcock brothers inc  21100.0       23250.0
264896                        big tree auto brokers   8625.0        8500.0
427777          bmw of annapolis/mini of annapolis    5475.0        3200.0
427829                        dt credit corporation   6200.0        4700.0
147329  ahfc/honda lease/trust/hvt  inc. eot acuras   25700.0       24500.0
…                                                 …      …            …
65502                         vw credit canada  inc   10500.0        7600.0
65518                         vw credit canada  inc   10100.0        7800.0
65522                         vw credit canada  inc    8600.0        8400.0
65526                         vw credit canada  inc   10600.0       10800.0
27667                    era classic auto sales ltd    1425.0        2150.0

                                       saledate
315386  Fri Feb 13 2015 01:00:00 GMT-0800 (PST)
264896  Fri Feb 06 2015 01:00:00 GMT-0800 (PST)
427777  Fri May 22 2015 02:15:00 GMT-0700 (PDT)
427829  Wed May 27 2015 03:00:00 GMT-0700 (PDT)
147329  Fri Jan 23 2015 01:00:00 GMT-0800 (PST)
…                                          …
65502   Tue Jan 06 2015 02:00:00 GMT-0800 (PST)
65518   Tue Jan 06 2015 02:00:00 GMT-0800 (PST)
65522   Tue Jan 06 2015 02:00:00 GMT-0800 (PST)
65526   Tue Jan 06 2015 02:00:00 GMT-0800 (PST)
27667   Thu Dec 18 2014 18:30:00 GMT-0800 (PST)

[63379 rows x 15 columns]
```

[15]:
```
transmission_mapping = df.dropna(subset=["transmission"]).groupby(["make",␣
↪"model", "trim"])["transmission"].first().to_dict()

# Update missing transmission values using the mapping
mask = df["transmission"].isna()
keys = df.loc[mask, ["make", "model", "trim"]].apply(tuple, axis=1)
df.loc[mask, "transmission"] = keys.map(transmission_mapping)
```

- Let's drop the rest of missing records

[16]:
```
df.dropna(subset=["transmission"], inplace=True)
```

3 Adjusting model column

[17]:
```
df['model'].unique()
```

[17]:
```
array(['Sorento', '3 Series', 'S60', '6 Series Gran Coupe', 'Altima',
       'M5', 'Cruze', 'A4', 'Camaro', 'A6', 'Optima', 'Fusion', 'Sonata',
       'Q5', '6 Series', 'Impala', '5 Series', 'A3', 'XC70', 'X5', 'SQ5',
       'S5', 'Verano', 'Suburban', 'ELR', 'V60', 'X6', 'ILX', 'K900',
       'Malibu', 'RX 350', 'Versa', 'Elantra', 'Versa Note', 'A8', 'X1',
       'Enclave', 'TTS', '4 Series', 'Silverado 2500HD', 'MDX',
       'Silverado 1500', 'SRX', 'G Coupe', 'G Sedan', 'FX', 'Santa Fe',
       'Genesis', 'Equus', 'Sonata Hybrid', 'Accent', 'Veloster',
       'Elantra Coupe', 'Azera', 'Tucson', 'Genesis Coupe', 'Wrangler',
       'S-Class', 'GS 350', 'Outlander', 'C-Class', 'Mazda2', 'Rio', 'M',
       '370Z', 'Soul', 'Outlander Sport', 'SLK-Class', 'ES 350',
       'E-Class', 'Mazda3', 'Cooper Clubman', 'Cooper', 'CX-9', 'Forte',
       'Compass', 'JX', 'RX 450h', 'LR4', 'Mazda5', 'Range Rover Evoque',
       'LS 460', 'GLK-Class', 'Sportage', 'Grand Cherokee', 'MKX', 'mkt',
       'XF', 'GL-Class', 'M-Class', 'Cooper Countryman', 'Lancer',
       'Range Rover Sport', 'Passat', 'Corolla', 'XC60', 'Sienna', 'Juke',
       'Yaris', 'Sentra', 'Rogue', 'NV', 'CC', 'Leaf', 'Camry', 'Tacoma',
       'Jetta', 'Impreza WRX', 'FJ Cruiser', 'Beetle', 'Avalon', 'FR-S',
       'NV200', 'RAV4', 'Quest', 'Tundra', 'tC', 'Maxima', 'Cayenne',
       '911', 'Xterra', 'Prius', 'S80', 'Frontier', 'Boxster',
       'Camry Hybrid', 'xB', 'Cube', 'Jetta SportWagen', '4Runner',
       'Sequoia', 'Legacy', 'Armada', 'Venza', 'Murano', 'Pathfinder',
       'Panamera', 'Forester', 'Highlander', 'Impreza', '750i', 'TSX',
       '7 Series', '1 Series', 'TL', '750li', 'S4', 'A7', 'A5', 'RDX',
       'M3', 'Cooper Coupe', 'ZDX', 'R8', 'X3', 'Avenger',
       'E-Series Wagon', 'Escape', 'Edge', 'Focus', 'Flex', 'Z4',
       'Traverse', 'F-350 Super Duty', 'Fiesta', '500', '200', 'Journey',
       'Charger', 'e350', 'Equinox', '300', 'F-150', 'Explorer',
       'Captiva Sport', 'Escalade', 'Grand Caravan', 'CTS Coupe',
       'Town and Country', 'E-Series Van', 'Volt', 'Express Cargo',
       'e150', 'X5 M', 'Expedition', 'Colorado', 'Express', 'California',
       'Escalade ESV', 'Sonic', 'Accord', 'CR-V', 'Mustang', 'Civic',
       'Fit', 'Pilot', 'Odyssey', 'Crosstour', 'Transit Connect',
       'Terrain', 'Taurus', 'G Convertible', 'Yukon', 'Veracruz', 'XJ',
       'Liberty', 'IS 250', 'XK', 'QX', 'CT 200h', 'Mazda6', 'MKZ',
       'Navigator', 'Range Rover', 'SL-Class', 'Sedona', 'IS 350',
       'Patriot', 'galant', '1500', 'GT-R', '2500', 'Galant', 'fortwo',
       'GLI', '5 Series Gran Turismo', 'XC90', 'Tiguan', 'GTI', 'Q7',
       'Highlander Hybrid', 'Prius Plug-in', 'CR-Z', 'EX', 'Sierra 1500',
       'LaCrosse', 'HHR', 'Accord Crosstour', 'CTS', 'Nitro', 'Tahoe',
       'Challenger', 'CTS-V', 'Escape Hybrid', 'X6 M', 'Ranger',
       'Insight', 'Fusion Hybrid', 'CTS-V Coupe', 'F-250 Super Duty',
       'Acadia', 'Impala Limited', 'Dart', 'Spark', 'M37', 'Sprinter',
```

'Town Car', 'CX-7', 'MKT', 'QX56', 'Aveo', 'Outback', 'Caliber',
'Routan', 'g1500', 'Sebring', 'Corvette', 'Continental GT Speed',
'malibu', 'Land Cruiser', 'town', 'V50', 'Commander',
'Altima Hybrid', 'G37 Convertible', 'g6', 'New Beetle', 'Golf',
'LR2', 'Lancer Sportback', 'G5', 'Yukon XL', 'Escalade Hybrid',
'Avalanche', 'Titan', 'Spectra', 'Rondo', 'Borrego', 'G-Class',
'MKS', 'CLK-Class', 'Tahoe Hybrid', 'Econoline Cargo',
'Econoline Wagon', 'PT Cruiser', 'CLS-Class', 'STS', 'Ridgeline',
'F-450 Super Duty', 'Magnum', 'Durango', 'S40', 'Malibu Classic',
'TT', 'Taurus X', 'Explorer Sport Trac', 'Ram Pickup 1500',
'impala', 'Cobalt', 'Pacifica', 'S6', 'Rabbit', 'C70',
'Sierra 2500HD', 'C30', 'VUE', 'GranTurismo', 'G6', 'Grand Prix',
'350Z', 'Raider', 'Mazdaspeed Mazda3', 'Solstice', 'Milan',
'GX 470', 'Aura', 'RX 400h', 'Matrix', 'H3', 'CL-Class', 'Outlook',
'7', 'G37', 'IS F', 'Touareg 2', 'Lancer Evolution', 'G35', 'xD',
'XJ-Series', 'G8', 'hhr', 'H2', 'DTS', 'lr3', 'sts',
'Silverado 1500 Classic', 'M45', 'Uplander', 'GS 450h',
'rangerover', 'Rendezvous', 'Monte Carlo', 'FX35', 'range', 'ION',
'R-Class', 'lancer', 'Eclipse', 'cx-7', 'B9 Tribeca', 'tundra',
'RSX', 'mazda5', 'Mariner', 'gx', 'Five Hundred', 'Envoy XL',
'S-Type', 'Element', 'Continental Flying Spur', 'S2000', 'FX45',
'sr', 'pilot', 'GS 430', 'Cayman S', 'Mark LT', 'ES 330', 'GS 300',
'Camry Solara', 'Touareg', 'Relay', 'lx', 'allroad quattro', '9-3',
'500L', 'C-Max Hybrid', 'pacifica', 'Freestyle', 'Ram Pickup 3500',
'Sprinter Cargo', 'DeVille', 'H2 SUT', 'TrailBlazer', 'Canyon',
'Dakota', 'Continental GT', 'Neon', 'Stratus', 'srx', 'Q45',
'Freestar', 'Montana', 'XLR', 'Aviator', 'g55', 'MPV', 'LS 430',
'Verona', 'Forenza', 'RX 330', '300M', 'SC 430', 'discovery',
'Excursion', 'Envoy XUV', 'Envoy', 'Concorde', 'Monterey',
'stratus', 'Mountaineer', 'Amanti', 'Malibu Maxx', 'Celica',
'Grand Am', 'Endeavor', 'Marauder', 'escape', 'QX4', 'LS',
'Blazer', 'Ram Pickup 2500', 'LeSabre', 'V40',
'Mazdaspeed Protege', 'Montero', 'ES 300', 'focus', 'Thunderbird',
'Century', '350z', 'Cavalier', 'Venture', 'S-10', 's55', 'Cougar',
'XL-7', 'Windstar', 'Silverado 1500HD', 'Explorer Sport',
'Savana Cargo', 'X-Type', 'Sonoma', 'IS 300', 'forester',
'Protege5', 'sprinter', 'RL', 'Alero', 'Grand Vitara', 'RX 300',
'L-Series', 'V70', 'Intrigue', 'XC', 'Discovery Series II',
'S-Series', 'alero', 'santa', 'ECHO', 'MX-5 Miata', 'Continental',
'Seville', 'camry', 'Park Avenue', 'Millenia', 'I30', 'gr',
'sienna', 'Cherokee', 'Z3', 'civic', 'ram', 'odyssey', 'taurus',
'expedition', 'Prizm', 'Escort', 's10', 'LHS', 'f250', 'Regal',
'explorer', 'G20', 'Bonneville', 'Eldorado', 'voyager', 'venture',
'durango', 'Intrepid', 'Contour', 'S90', 'Sunfire', 'mpv', '200SX',
'Rodeo', 'wrangler', 'caravan', 'f150', 'windstar', 'Tercel',
'S70', 'Discovery', 'Mustang SVT Cobra', 'pickup', 'Grand Marquis',
'ciera', 'Legend', 'LS 400', 'Cutlass Ciera', 'Santa Fe Sport',

'Cadenza', 'Q50', 'Elantra GT', 'F-TYPE', 'Shelby GT500', 'QX70',
'QX60', 'Q60 Convertible', 'Cooper Roadster', 'CX-5',
'Cooper Paceman', 'Rogue Select', 'Cayman', 'CLA-Class', 'allroad',
'ATS', 'Prius v', 'Continental GTC', 'XV Crosstrek', '3500',
'C-Max Energi', 'Focus ST', 'RS 7', 'GX 460', 'CTS Wagon',
'SLS AMG', 'Aspen', 'Eclipse Spyder', 'Vibe', 'Eos', 'Entourage',
'expeditn', 'Caravan', 'Quattroporte', 'M35', '9-5', 'SSR',
'Astro Cargo', 'Safari Cargo', 'passat', 'Tribute', 'Diamante',
'Sable', 'Silverado 3500', 'Phaeton', 'R32', 'I35', 'Bravada',
'Tahoe Limited/Z71', 'Truck', 'C/K 1500 Series', 'grand', 'SC 300',
'Roadmaster', 'SC 400', 'LX 570', 'QX80', 'RS 5', 'Jetta GLI',
'ES 300h', 'capt', 'M4', 'SX4', 'iQ', 'Kizashi', 'C/V Cargo Van',
'Prius c', '750lxi', 'alp', 'Lucerne', 'Escalade EXT',
'Silverado 3500HD', 'Crown Victoria', 'Sierra 3500HD', 'M56',
'IS 250 C', '3', 'endeavor', 'corolla', 'ActiveHybrid X6', 'dts',
'g3500', 'colorado', 'sebring', 'e250', 'police',
'Elantra Touring', 'G37 Coupe', 'HS 250h', 'journey',
'Mazdaspeed3', 'Milan Hybrid', 'Ghost', 'Silverado 1500 Hybrid',
'Yukon Hybrid', 'optima', 'borrego', 'mazda6', '6',
'Mariner Hybrid', 'Torrent', 'VUE Hybrid', 'G3', '9-7X', 'vibe',
'Expedition EL', 'Tiburon', 'patriot', 'LR3', 'Navigator L',
'Astra', 'Tribeca', 'XL7', 'sx4', 'Sky', 'Reno', 'M6', 'S8',
'Terraza', 'Silverado 2500HD Classic', 'corvette', 'ram3500',
'Sierra 1500 Classic', 'Sierra 2500HD Classic', 'Fusion Energi',
'XK-Series', 'quattroporte', 'B-Series Truck', 'mazda3',
'Mazdaspeed Mazda6', 'Montego', 'g5', 'Rainier', 'TrailBlazer EXT',
'Crossfire', 'magnum', 'savana', 'Sierra 1500HD', 'ridgelin',
'Savana', 'Zephyr', 'rrs', 'tribute', 'Montana SV6', 'tt',
'Classic', 'pt', 'freestyle', 'Ascender', 'XG350', 'Q60 Coupe',
'Q70', 'QX50', 'GTO', 'xA', 'L300', 'Baja', '9-2X', 'Aerio',
'rainier', 'Silverado 2500', 'Astro', 'Tracker', 'intrepid',
'F-150 Heritage', 'expedit', 'accord', 'Freelander', 'MR2 Spyder',
'RS 6', 'Voyager', 'Axiom', 'cl55', 'Montero Sport', 'sl55',
'Protege', 'Silhouette', 'b1500', 'concorde', '626', 'Blackwood',
'Rodeo Sport', 'LX 470', 'Villager', 'Firebird', 'Aurora', 'CL',
'EuroVan', 'Catera', 'Leganza', 'XG300', 'Prelude', 'Trooper',
'Prowler', 'Aztek', 'Cabrio', 'Integra', 'cavalier', 'astro',
'excurs', 'dakota', 'Cirrus', 'crown', 'twn&country', 'Jimmy',
'safari', 'ranger', 'sonoma', 'Sierra 2500', 'yukon', 'Sephia',
'Passport', 'Mirage', 'silhouette', 'villager', 'beetle',
'suburban', 'Lumina', 'Amigo', 'mountaineer', 'Esteem',
'pathfinder', 'quest', 'Breeze', 'montana', 'intrigue', 'lumina',
'envoy', 'Cutlass', 'GS 400', 'E-150', 'camaro', 'Regency',
'thunderbird', 'B-Series Pickup', 'legacy', 'bronco', 'Le Baron',
'Caprice', 'Pickup', 'century', '500-Class', '300-Class', 'previa',
'Murano CrossCabriolet', 'NV Cargo', 'Jetta Hybrid', 'S7',
'Encore', 'XTS', 'Black Diamond Avalanche', 'RLX', '2 Series',

```
'M6 Gran Coupe', 'regal', 'BRZ', 'C/V Tradesman', 'Model S',
'Beetle Convertible', 'Golf R', 'TSX Sport Wagon', 'interstate',
'Corvette Stingray', 'a6', 'SS', 'Mazdaspeed 3', 'i-MiEV', 'F430',
'EX35', 'yaris', 'GranSport', 'escalade', 'f350', 'Yukon Denali',
'Ghibli', '960', 'Cutlass Supreme', 'cougar', 'IS 350 C', 'a4',
'cobalt', 'uplander', 'Mazdaspeed MX-5 Miata', 'Ram Cargo',
'Safari', 'Eighty-Eight', '850', 'J30', 'Promaster Cargo Van',
'rr', '1', 'Malibu Hybrid', '350', 'twn/cntry', 'Spirit',
'Accord Hybrid', '3 Series Gran Turismo', 'e', 'crossfire',
'Viper', 'Riviera', 'Avalon Hybrid', 'routan', 'RX-8',
'V8 Vantage', 'Equator', 'sportage', 'C/K 3500 Series',
'Mark VIII', 'charger', 'GranTurismo Convertible', 'avenger',
'equinox', 'Vitara', 'avalon', 'siera', 'pathfind',
'Eighty-Eight Royale', 'cherokee', 'ActiveHybrid 7', 'GS 460',
'Tribute Hybrid', 'Aura Hybrid', 'tahoe', 'eurovan', 'g2500',
'Golf GTI', 'i-Series', 'MKZ Hybrid', 'Macan', 'FX50', 'comm',
'STS-V', 'Windstar Cargo', 'jetta', 'CTS-V Wagon', 'Karma', 'Z4 M',
'matrix', 'subrbn', 'b2300', 'mountnr', 'Coupe', 'uplandr',
'Ram Van', 'Tempo', 'Tracer', 'CV Tradesman', 'DB9',
'C/K 2500 Series', 'i8', 'Rapide', 'Nubira', 'Corsica',
'NV Passenger', 'Spyder', 'LS 600h L', '400-Class', 'H3T',
'LX 450', 'WRX', 'Silverado 3500 Classic', '500e',
'Continental Supersports', 'Sierra 3500', 'Mystique',
'F-150 SVT Lightning', '190-Class', 'MKC', 'Aspire', '940',
'Gallardo', 'Continental Flying Spur Speed', '3000GT', 'TT RS',
'300ZX', 'ActiveHybrid 5', 'Sierra 1500 Hybrid', 'ML55 AMG',
'S-10 Blazer', 'RS 4', 'T100', 'Continental GTC Speed', 'mdx',
'Transit Van', 'Sidekick', 'E-250', '8 Series', '420-Class',
'E-350', 'Achieva', 'B-Class Electric Drive', 'Paseo',
'Civic del Sol', 'Exige', 'X4', 'Spark EV', 'Transit Wagon', 'H1',
'SLS AMG GT', 'Metro', 'Grand Cherokee SRT', 'RC F', 'Q3',
'4 Series Gran Coupe', 'RC 350', '360', 'GLA-Class', 'TLX',
'458 Italia'], dtype=object)
```

[18]:
```
df['model'] = df['model'].apply(lambda x: x.capitalize() if x.islower() else x)
```

4 Adjusting state column

[19]:
```
df['state'].unique()
```

[19]:
```
array(['ca', 'tx', 'pa', 'mn', 'az', 'wi', 'tn', 'md', 'fl', 'ne', 'nj',
       'nv', 'oh', 'mi', 'ga', 'va', 'sc', 'nc', 'in', 'il', 'co', 'ut',
       'mo', 'ny', 'ma', 'pr', 'or', 'la', 'wa', 'hi', 'qc', 'ab', 'on',
       'ok', 'ms', 'nm', 'al', 'ns'], dtype=object)
```

[20]:
```
df = df[~df['state'].isin(['3vwd17aj4fm201708', 'ns',
       '3vwd17aj2fm258506', '3vwd17aj3fm276741', '3vwd17aj2fm285365',
       '3vwd17aj0fm227318', '3vwd17aj6fm218641', '3vwd17aj7fm223475',
       '3vwd17aj5fm297123', '3vwd17aj5fm219943', '3vwd17aj9fm219766',
       '3vwd17aj3fm259017', '3vwd17aj5fm206111', '3vwd17aj5fm273601',
       '3vwd17aj5fm221322', '3vwd17aj5fm268964', '3vwd17aj6fm231972',
       '3vwd17aj7fm222388', '3vwd17aj7fm218440', '3vwd17ajxfm315938',
       '3vwd17aj7fm229552', '3vwd17aj8fm298895', '3vwd17aj4fm236636',
       '3vwd17aj5fm225953', '3vwd17aj7fm326640', '3vwd17aj8fm239622',
       '3vwd17aj2fm261566'])]
```

[21]:
```
df['state'] = df['state'].str.upper()
```

5 Adjusting make column

[22]:
```
df['make'].unique()
```

[22]:
```
array(['Kia', 'BMW', 'Volvo', 'Nissan', 'Chevrolet', 'Audi', 'Ford',
       'Hyundai', 'Buick', 'Cadillac', 'Acura', 'Lexus', 'Infiniti',
       'Jeep', 'Mercedes-Benz', 'Mitsubishi', 'Mazda', 'MINI',
       'Land Rover', 'Lincoln', 'lincoln', 'Jaguar', 'Volkswagen',
       'Toyota', 'Subaru', 'Scion', 'Porsche', 'bmw', 'Dodge', 'FIAT',
       'Chrysler', 'ford', 'Ferrari', 'Honda', 'GMC', 'mitsubishi', 'Ram',
       'smart', 'chevrolet', 'Bentley', 'chrysler', 'pontiac', 'Pontiac',
       'Saturn', 'Maserati', 'Mercury', 'HUMMER', 'landrover', 'cadillac',
       'land rover', 'mazda', 'toyota', 'lexus', 'gmc truck', 'honda',
       'porsche', 'Saab', 'mercedes', 'Suzuki', 'dodge', 'nissan',
       'subaru', 'Oldsmobile', 'oldsmobile', 'hyundai', 'jeep', 'Isuzu',
       'dodge tk', 'Geo', 'volkswagen', 'suzuki', 'Rolls-Royce', 'kia',
       'gmc', 'maserati', 'mercury', 'audi', 'buick', 'mercedes-b',
       'Daewoo', 'chev truck', 'ford tk', 'plymouth', 'vw', 'Plymouth',
       'ford truck', 'Tesla', 'airstream', 'Aston Martin', 'Fisker',
       'Lamborghini', 'Lotus'], dtype=object)
```

[23]:
```
df["make"] = df['make'].str.capitalize()
df.loc[(df["make"] == "Vw"),"make"] = 'Volkswagen'
df.loc[(df["make"] == 'Mercedes-b'),"make"] = 'Mercedes-benz'
df.loc[(df["make"] == 'Ford tk'),"make"] = 'Ford truck'
df.loc[(df["make"] == 'Hyundai tk'),"make"] = 'Hyundai truck'
df.loc[(df["make"] == 'Dodge tk'),"make"] = 'Dodge truck'
```

6 Adjusting body column

[24]:
```
df['body'].unique()
```

[24]:
```
array(['SUV', 'Sedan', 'Convertible', 'Coupe', 'Wagon', 'Hatchback',
       'Crew Cab', 'G Coupe', 'G Sedan', 'Elantra Coupe', 'Genesis Coupe',
       'Minivan', nan, 'Van', 'Double Cab', 'CrewMax Cab', 'Access Cab',
       'King Cab', 'SuperCrew', 'CTS Coupe', 'Extended Cab',
       'E-Series Van', 'SuperCab', 'Regular Cab', 'G Convertible', 'Koup',
       'Quad Cab', 'CTS-V Coupe', 'sedan', 'G37 Convertible', 'Club Cab',
       'Xtracab', 'Q60 Convertible', 'CTS Wagon', 'convertible',
       'G37 Coupe', 'Mega Cab', 'Cab Plus 4', 'Q60 Coupe', 'Cab Plus',
       'Beetle Convertible', 'TSX Sport Wagon', 'Promaster Cargo Van',
       'GranTurismo Convertible', 'CTS-V Wagon', 'Ram Van', 'minivan',
       'suv', 'Transit Van', 'van', 'regular-cab', 'g sedan', 'g coupe',
       'hatchback', 'king cab', 'supercrew', 'g convertible', 'coupe',
       'crew cab', 'wagon', 'double cab', 'e-series van', 'regular cab',
       'quad cab', 'g37 convertible', 'supercab', 'extended cab',
       'crewmax cab', 'genesis coupe', 'access cab', 'mega cab',
       'xtracab', 'beetle convertible', 'cts coupe', 'koup', 'club cab',
       'elantra coupe', 'q60 coupe', 'cts-v coupe', 'transit van',
       'granturismo convertible', 'tsx sport wagon',
       'promaster cargo van', 'q60 convertible', 'g37 coupe',
       'cab plus 4', 'cts wagon'], dtype=object)
```

[25]:
```
df.loc[(df['body'].str.contains("cab"))  & (~df['body'].isna()) ,'body'].
↪unique()
```

[25]:
```
array(['Xtracab', 'regular-cab', 'king cab', 'crew cab', 'double cab',
       'regular cab', 'quad cab', 'supercab', 'extended cab',
       'crewmax cab', 'access cab', 'mega cab', 'xtracab', 'club cab',
       'cab plus 4'], dtype=object)
```

[26]:
```
df['body'] = df['body'].str.capitalize()
df.loc[(df["body"] == 'Regular-cab'),"body"] = 'Regular cab'
```

[27]:
```
df[df['body'].isna()]
```

[27]:

	year	make	model	trim	body	transmission	state
468	2013	Lincoln	Mkt	awd v6	NaN	Automatic	CA
743	2012	Bmw	750i	xdr 750i xdriv	NaN	Automatic	CA
770	2012	Bmw	750li	750li	NaN	Automatic	CA
793	2012	Bmw	750i	750i	NaN	Automatic	CA
794	2012	Bmw	750li	750li	NaN	Automatic	CA
…	…	…	…	…	…	…	…
143161	2007	Ford	Expeditn	el 4x4 limited	NaN	Automatic	FL
143419	2005	Cadillac	Srx	awd v6 awd	NaN	Automatic	GA
145012	2009	Kia	Borrego	lx	NaN	Automatic	IL
145861	2005	Ford	Explorer	4x4 v6 xlt	NaN	Automatic	NY
146156	2003	Lexus	Gx	470	NaN	Automatic	GA

```
        condition  odometer      color interior  \
468          41.0   74874.0      black    black
743           4.0   50790.0       gray    black
770          37.0   31762.0      black    black
793          49.0   53016.0      white     gray
794          34.0   24739.0      white     gray
…              …         …          …        …
143161       35.0   99631.0      brown    black
143419       24.0   76004.0  off-white    beige
145012       29.0  112614.0       gray    black
145861       28.0   84017.0     silver     gray
146156       39.0  167381.0     silver     gray

                                                seller      mmr  sellingprice  \
468     remarketing by ge/manheim southern california  19300.0       17750.0
743            financial services remarketing (lease)  33900.0       33500.0
770            financial services remarketing (lease)  45000.0       45000.0
793            financial services remarketing (lease)  31300.0       34250.0
794            financial services remarketing (lease)  47000.0       46750.0
…                                                 …       …            …
143161                                 mint motors  12600.0       13200.0
143419                        germain bmw of naples   6350.0        6600.0
145012             remarketing by ge/manheim chicago   7300.0        7000.0
145861                          adk auto brokers inc   5125.0        4900.0
146156                           innovative auto llc   9375.0       10000.0

                                     saledate
468     Thu Dec 18 2014 12:00:00 GMT-0800 (PST)
743     Thu Dec 18 2014 12:30:00 GMT-0800 (PST)
770     Thu Dec 18 2014 12:30:00 GMT-0800 (PST)
793     Thu Dec 18 2014 12:30:00 GMT-0800 (PST)
794     Thu Dec 18 2014 12:30:00 GMT-0800 (PST)
…                                          …
143161  Wed Jan 14 2015 09:00:00 GMT-0800 (PST)
143419  Thu Jan 15 2015 02:00:00 GMT-0800 (PST)
145012  Thu Jan 15 2015 02:00:00 GMT-0800 (PST)
145861  Thu Jan 15 2015 01:45:00 GMT-0800 (PST)
146156  Thu Jan 15 2015 02:00:00 GMT-0800 (PST)

[2396 rows x 15 columns]
```

[28]:
```
body_mapping = df.dropna(subset=["body"]).groupby(["make", "model"])["body"].
  ↪first().to_dict()

# Update missing body values using the mapping
mask = df["body"].isna()
keys = df.loc[mask, ["make", "model"]].apply(tuple, axis=1)
```

```
df.loc[mask, "body"] = keys.map(body_mapping)
```

[29]:
```
df.dropna(subset=["body"],inplace = True)
```

[30]:
```
df.isna().sum()
```

[30]:
```
year                0
make                0
model               0
trim                0
body                0
transmission        0
state               0
condition       11286
odometer            0
color               0
interior            0
seller              0
mmr                12
sellingprice       12
saledate           12
dtype: int64
```

7 Adjusting trim column

[31]:
```
list(df['trim'].unique())
```

[31]:
```
['LX',
 '328i SULEV',
 'T5',
 '650i',
 '2.5 S',
 'Base',
 '1LT',
 '2.0T Premium Plus quattro',
 'LT',
 '3.0T Prestige quattro',
 'SE',
 '2LT',
 'LS',
 'LTZ',
 '528i',
 '1.8 TFSI Premium',
 'T6',
 'sDrive35i',
 '3.0T Premium Plus quattro',
```

```
'Premium Plus quattro',
'Convenience Group',
'xDrive35i',
'Technology Package',
'Luxury',
'1.6 SL',
'1.6 SV',
'L 3.0T quattro',
'sDrive28i',
'Leather Group',
'quattro',
'428i SULEV',
'Work Truck',
'Advance and Entertainment Packages',
'535d',
'G37 Sport',
'G37 Journey',
'FX37',
'Limited',
'5.0 R-Spec',
'Signature',
'GLS',
'Sport',
'Sport 2.0T',
'Turbo',
'GS',
'Ultimate',
'3.8',
'3.8 Track',
'Unlimited Rubicon',
'EX Hybrid',
'2.5',
'S550',
'C250',
'SX',
'Unlimited Sahara',
'C250 Sport',
'Touring',
'M37',
'+',
'LE',
'SLK350',
'E350 Sport',
'i SV',
'Latitude',
'i Touring',
'JX35',
```

```
'G37x',
'HSE LUX',
'EX',
'i Sport',
'Pure Plus',
'L',
'GLK350',
'Unlimited Sport',
'E350 Sport BlueTEC',
'Pure',
'GL450 4MATIC',
'ML350',
'Pure Premium',
'Ralliart',
'C63 AMG',
'Supercharged Limited Edition',
'E350',
'ES',
'SE PZEV',
'3.2',
'LE 7-Passenger Mobility Auto Access',
'S',
'SR',
'1500 S',
'R-Line PZEV',
'SL',
'XLE',
'PreRunner',
'3.5 SL',
'1.6 S Plus',
'2.5L PZEV',
'Wolfsburg Edition PZEV',
'TDI',
'Tundra',
'3.5 SV',
'Carrera S',
'Two',
'SV',
'1.8 SL',
'S PZEV',
'FE+ S',
'SR5',
'3.5 S',
'2.5i Premium PZEV',
'PreRunner V6',
'S Hybrid',
'S Special Edition',
```

```
'2.5X PZEV',
'2.0i Sport Limited PZEV',
'V6',
'STI',
'FE+ SV',
'TDI SE',
'3.0T Premium quattro',
'740i',
'335is',
'328i',
'135i',
'640i',
'535i',
'740Li',
'528i xDrive',
'Premium quattro',
'335i',
'2.0T Premium quattro',
'128i',
'Special Edition',
'550i',
'650i xDrive',
'4.2 quattro',
'xDrive28i',
'3.2 Premium Plus quattro',
'Technology and Entertainment Packages',
'535i xDrive',
'E-350 Super Duty XL',
'SEL',
'xDrive35i Premium',
'Titanium',
'xDrive35d',
'LT Fleet',
'LS Fleet',
'Lariat',
'SES',
'XLS',
'SXT',
'SRT8',
'xDrive50i',
'S V6',
'XLT',
'C Pop',
'LTZ Fleet',
'American Value Package',
'Premium',
'FX2',
```

```
'E-150',
'Hybrid',
'1500',
'Eco',
'2LS Fleet',
'E-350 Super Duty',
'1LT Fleet',
'LS 3500',
'XL',
'Crew',
'Platinum Edition',
'G25',
'LX-P',
'LX-S',
'4.6',
'EX V-6',
'EX-L',
'EX-L V-6',
'GT',
'Cargo Van XLT',
'SLE-2',
'G37',
'FX35',
'Denali',
'HF',
'E350 Luxury',
'XKR',
'QX56',
'XJL Supercharged',
'i Grand Touring',
'Laredo',
'M35h',
'Portfolio',
'HSE',
'S ALL4',
'Koup EX',
'SL550',
's Touring',
'Supercharged',
'M37x',
'fe',
'2',
'1.8 S',
'R/T',
'ST',
'3500 S',
'WRX',
```

```
'Black Edition',
'ML350 4MATIC',
'2.5X Touring PZEV',
'SLT',
'S350 BlueTEC 4MATIC',
'Carrera S (Midyear Redesign)',
'4',
'E350 Luxury 4MATIC',
'ML350 BlueTEC',
'2.5i Premium',
'1.6 S',
'passion coupe',
'2.0T Premium',
'Autobahn PZEV',
'Three',
'SH-AWD',
'328i xDrive',
'PZEV',
'Sport PZEV',
'3.0T S line Prestige quattro',
'2.0T White Turbo Launch Edition PZEV',
'TDI Premium quattro',
'Lux Limited PZEV',
'335d',
'Base 7-Passenger',
'335i xDrive',
'LT1',
'C',
'Express',
'FX4',
'EX35',
'LTZ 1500',
'SLE',
'CXL',
'STX',
'CXS',
'SE Fleet',
'CX',
'Panel LS',
'GLS PZEV',
'SPORT',
'LT2',
'2500',
'E-250',
'Heat',
'SLE-1',
'Si',
```

```
'Touring-L',
'Aero',
'!',
's Sport',
'G25x',
'GL550 4MATIC',
'2500 144 WB Passenger',
'IPL',
'S400 Hybrid',
'Signature Limited',
'XJL',
'C350 Sport',
'E350 BlueTEC',
'C300 Sport',
'E350 4MATIC',
's Grand Touring',
'2.0 SR',
'2.0 S',
'XLE 8-Passenger',
'SEL PZEV',
'Platinum',
'2.5i',
'SXT Fleet',
'2.0 SL',
'1SS',
'750Li',
'1500 LT',
'3.5L',
'2500 170 WB Passenger',
'ls',
'3.6 Premium quattro',
'Performance',
'T5 R-Design',
'X',
'EL Limited',
'3.5 SR',
'Tundra Grade',
'S600',
'4c',
'Blue',
'2.5L Red Rock Edition PZEV',
'3.2 R-Design',
'2.0T quattro',
'Harley-Davidson',
'III',
'DX-VP',
'GLK350 4MATIC',
```

```
'T6 R-Design',
'GTS',
'S63 AMG',
'BRABUS coupe',
'GL550',
'GL450',
'Grand Touring',
'ES-Sport',
'Premium Luxury',
'xDrive48i',
'DE',
'SS',
'SportWagen S PZEV',
'XLE 7-Passenger',
'xDrive30i',
'1500 LTZ',
'G550',
'SLT-2',
'CLK550',
'TDI Loyal Edition',
'C300 Luxury',
'SE V6',
'CLS550',
'GL320 BlueTEC',
'V8',
'RT',
'3.0si',
'2.0T',
'2.4i',
'3.6 quattro',
'335xi',
'3.2 quattro',
'L quattro',
'SLT-1',
'v6 ltz',
'3.0i',
'4.8i',
'1500 LS',
'328xi',
'XR',
'Type-S',
'WRX STI',
'S550 4MATIC',
'525i',
'Premier',
'G500',
'XE',
```

```
'3.2 Special Edition',
'GXP',
'RTS',
'CL550',
'Journey',
'530i',
'VR6 FSI',
'GSR',
'LE 7-Passenger',
'Unlimited X',
'VR6',
'XJ8 L',
'SE SULEV',
'King Ranch',
'ZX3 S',
'Deluxe',
'ZX4 SE',
'LS2',
'Cargo',
'Overland',
'LT3',
'SEL Fleet',
'SLE 1500',
'R350',
'CLK350',
'es',
'325i',
'2.5 PZEV',
'325xi',
'Limited 7-Passenger',
'CE',
'1.8T',
'Nismo',
'330i',
'4x2 v8 sr5',
'WRX Limited',
'3.5 SE',
'530xi',
'touring',
'Wolfsburg Edition',
'Value Leader',
'Base PZEV',
'SL 1500',
'ZX4 SES',
'SLE1',
'3500',
'3',
```

```
'Eddie Bauer',
'Laramie',
'LS 1500',
'EX-P',
'Standard',
'4x4 ex-l',
'Luxury III',
'SRT-8',
'SL2',
'LE 8-Passenger',
'XLE Limited 7-Passenger',
'CLK500',
'C230 Sport',
'2.5 i',
'XLE V6',
'CLS500',
'CL500',
'ML500',
'E55 AMG',
'E500',
'SLE V6',
'4x2 6c carrera s',
'Value Edition PZEV',
'Easy',
'E-350 Super Duty XLT',
'Lounge',
'Pop',
'EL King Ranch',
'fwd v6 touring',
'545i',
'awd v6',
'2500 High Roof 140 WB',
'645Ci',
'SRT-10',
'Z85 SLE',
'1500 SLE',
'GL',
'4.4i',
'Fleet',
'745Li',
'Value Package',
'HEV',
'GT Premium',
'C230 Kompressor',
'E320',
'E500 4MATIC',
'Bi-Color Edition',
```

```
'Dark Flint Edition',
'SV6',
'SL55 AMG',
'CE 7-Passenger',
'745i',
'1.8T quattro',
'GLS 1.8T',
'WRX STi',
'2.5i Limited',
'i',
's',
'Executive L Fleet',
'C240 Luxury',
'S430',
'CLK320',
'2.5 RS',
'Enthusiast',
'LE V6',
'C230',
'awd carrera 4s',
'GT1',
'GTP',
'ZX3',
'Arc',
'SL500',
'CL600',
'Touring Edition',
'4c se',
'Convenience',
'NBX',
'Linear',
'SE1',
'awd',
'C320',
'v6 ls',
'LXi',
'SES Deluxe',
'2.5T',
'LPT',
'DX',
'4x4 v6 limited',
'3.0 quattro',
'XLT Value',
'GLX',
'GLS 1.8T 4Motion',
'325Ci',
'Custom',
```

```
'2500 SLE',
'330Ci',
'C240',
'2.4',
'zx3',
'XLT Popular',
'180hp',
'se',
'Z06',
'XLT Premium',
'LX Standard',
'ZX5',
'Choice',
'Sahara',
'2.5 SL',
'SLS',
'GLX VR6',
'awd xs',
'E320 4MATIC',
'2.4T',
'DHS',
'XJ8',
'SLK320',
'Avant quattro',
'3.5',
'W8 4Motion',
'GT Deluxe',
'GL1',
'L300',
'VDC',
'Cartier',
'GX',
'LW300',
'GXE',
'gl',
'4.6 HSE',
'2.8 quattro',
'S55 AMG',
'SD',
'ML430',
'SE-V6',
'ML320',
'v6 le',
'740iL',
'1.8T Avant quattro',
'VP',
'323Ci',
```

```
'GLE',
'XC',
'le',
'GLX V6',
'GLX 1.8T',
'Vanden Plas',
'Classic',
'2.3',
'dx',
'2.7T quattro',
'ex',
'lx',
'xle',
'4x2 xlt',
'323i',
'Executive',
'awd v8 limited',
'Ultra',
'540i',
'fwd 4c',
'fwd ext ls',
'4x4 slt',
'EX V6',
'SE Sport',
'S320 LWB',
'SL1',
'Outback Limited',
'GLT',
'JXi',
'awd xlt',
'S320 SWB',
'XJ8L',
'RS',
'4x4 4c se',
'fwd gl',
'SC1',
'328is',
'ce',
'4x4 dx',
'C280',
'3.8 Ultimate',
'V8 S',
'Forte5 EX',
'Hybrid Sport',
'XJR',
'Hybrid Premium',
'Touring w/Navigation and Rear Entertainment',
```

```
'SLT 1500',
'John Cooper Works ALL4',
'XXV',
'Autobiography LWB',
'Autobiography',
'Club',
'CLA45 AMG',
'GL63 AMG',
'E550',
'G63 AMG',
'C300 Sport 4MATIC',
'E63 AMG 4MATIC',
'GLK250 BlueTEC 4MATIC',
'CLA250',
'320i',
'Premium Package',
'Prestige',
'LE Plus',
'S Plus',
'2.0XT Touring',
'V6 Executive',
'750Li xDrive',
'LX PZEV',
'EX-L w/Rear Entertainment',
'Hybrid PZEV w/Leather and Navigation',
'Prestige quattro',
'SH-AWD w/Technology Package',
'L 4.2 quattro',
'Tradesman',
'C300 Luxury 4MATIC',
'550i xDrive',
'SLK300',
'3.6R Limited',
'SRT8 Core',
'Limited Edition PZEV',
'SportWagen TDI',
'C HEMI',
'R350 4MATIC',
'4x2 6c carrera',
'i Sport Value Edition',
'4x2 v8 slt',
'Limited Edition',
'SEL Plus',
'SLK280',
'S65 AMG',
'2.5i Special Edition',
'awd 6c carrera 4s',
```

```
'C280 Luxury',
'3.6',
'fwd v6',
'Spectra5',
'EX PZEV',
'S500',
'2.5 GT Limited',
'330xi',
'fwd 4c gls',
'2.7T S-Line quattro',
'ZQ8 LS',
'LX-V6',
'LX Family Value',
'awd ex',
'XE-V6',
'SRT-4',
'1.8',
'4x2 carrera',
'Linear 2.3T',
'GLX 4Motion',
'ES-V6',
'L100',
'DTS',
'SLK230',
'GLS TDI',
'ESC',
'ZX2',
'G3500',
'awd l',
'Laramie SLT',
'GLX VR6 (1999.5)',
'4.0 S',
'sport',
'cpe carrera 2',
'454SS',
'TDI Prestige quattro',
'Edition 30 PZEV',
'2.0T Premium Plus',
'3.8 R-Spec',
'2.0 TFSI Premium Plus quattro',
'2.5i PZEV',
'2.5X Premium',
'2.0i Premium PZEV',
'2.5i Limited PZEV',
'LE Popular Package',
'Laramie Longhorn Edition',
'2.5X Premium PZEV',
```

```
'2.0i',
'passion cabriolet',
'R-Line',
'2.5L',
'SE 8-Passenger',
'Lux Limited',
'SEL 4Motion',
'TDI Sport',
'TDI SEL Premium',
'2.0L TDI',
'2.0T Avant Premium quattro',
'2.0 TDI Premium',
'3.2 Premium',
'CXL1',
'CXL Turbo',
'CXL2',
'2SS',
'Aveo5 1LT',
'LT 1500',
'Mainstreet',
'Lux',
'Rush',
'C/V',
'v6 s',
'Citadel',
'LX Fleet',
'Shock',
'Lariat Limited',
'SVT Raptor',
'Cargo Van XL',
'SHO',
'3.8 Grand Touring',
'Limited PZEV',
'x',
'XFR',
'EX Turbo',
'Rubicon',
'GL350 BlueTEC',
'E63 AMG',
'E550 4MATIC',
'awd ls',
'R350 BlueTEC',
'ML550 4MATIC',
'PRO-4X',
'1.8 S Krom Edition',
'SE-R',
'BigHorn',
```

```
'1.6',
'SE-R Spec V',
'Outback Sport',
'2.5X',
'SportBack Technology',
'Turbo4 Premium',
'2.5X Touring',
'2.5X Limited',
'2.5XT Premium',
'Sport GTS',
'Sport SLS',
'Limited FFV',
'Lux SULEV',
'R-Design',
'Tundra FFV',
'750i',
'750i xDrive',
'3.2 Premium quattro',
'Super',
'V6 Luxury Sport',
'Aveo5 LS',
'police',
'4c ls w/1ls',
'4x2 rg work truck',
'4c ls w/1fl',
'Hero',
'RT Fleet',
'Limited Fleet',
'v6 touring',
'Touring Plus',
'EL SSV Fleet',
'Wagon XLT',
'RTL',
'Renegade',
'Limited V6',
'Unlimited Sport RHD',
'Koup SX',
'awd v6 r/t',
'E-150 XL',
'2500 144 WB Cargo',
's Touring Plus',
'1.8 Base',
'MR',
'MR Touring',
'4S',
'Carrera',
'Crossover',
```

```
'XSport',
'S Krom Edition',
'V6 TDI',
'3.0T quattro',
'Komfort',
'SportWagen SE',
'Komfort PZEV',
'2.5L Red Rock Edition',
'1500 LT1',
'Aveo5 2LT',
'SLT1',
'SLT2 1500',
'SLE2 XFE',
'GLS V6',
'Rio5 LX',
'ML320 BlueTEC',
'i sport',
'ML550',
'John Cooper Works',
'E320 BlueTEC',
'1.8 Krom',
'2.5XT Limited',
'4.2i',
'Turbo S',
'awd 4c',
'VR6 4Motion',
'SportWagen SE PZEV',
'2.0T PZEV',
'SEL Premium',
'528xi',
'535xi',
'4.2 Premium quattro',
'Aveo5 Special Value',
'V8 Luxury',
'2500 LT',
'v6 lt',
'Panel LT',
'4x4 v8 slt',
'3500 170 WB',
'2500 144 WB',
'60th Anniversary',
'FX4 Off-Road',
'Alpha',
'4x4 4c limited',
'3.0L',
'V8 SE',
'CL63 AMG',
```

```
'GL320 CDI',
'ML320 CDI',
'R320 CDI',
'S SULEV',
'SL SULEV',
'5.3i',
'Limited 5-Passenger',
'passion',
'3.0 R L.L.Bean Edition',
'2.5I',
'Red Line',
'2.5 X Premium Package',
'2.5I Limited',
'passion cabrio',
'V8 FSI',
'525xi',
'760Li',
'z06',
'SSV Fleet',
'ZX5 SES',
'ZX4 ST',
'Special Edition V-6',
'SC',
'XK',
'Rio5 SX',
'exec gt',
'B4000',
'CLS63 AMG',
'R500',
'E320 BLUETEC',
'C280 Luxury 4MATIC',
'gt',
'SE Off-Road',
'2.5 XT Limited',
'awd x premium',
'awd x',
'4x4 v8 sr5',
'4x2 v8 limited',
'Triple White',
'4x2 v6 sr5',
'GLI',
'Wagon Titanium LWB',
'police police',
'awd v6 touring',
'awd v6 limited',
'4x2 v6',
'4x2 v6 sxt',
```

```
'2500 High Roof 158 WB',
'2500 140 WB',
'4x4 4c hybrid',
'ZX5 SE',
'comm',
'LX Special Edition',
'LX V-6',
'Limited SULEV',
'XK8',
'Wagon XLT LWB',
'6c unlimited',
'6c rubicon',
'6c rubicon ul',
'E320 CDI',
'CL55 AMG',
'4wd i',
'C55 AMG',
'SLK55 AMG',
'IX',
'5-Passenger',
'Value Edition',
'3.6 4Motion',
'Z71',
'Z85 LS Base',
'Z85',
'ZQ8',
'LS Sport',
'LT Entertainer',
'v6 se',
'4x2 v6 se',
'v6 sxt',
'XLT Sport',
'XLT NBX',
'4x4 v6 xlt',
'4x2 v6 xlt',
'4x4 v6 ed bauer',
'xls',
'4x2 v6 limited',
'EDGE',
'ZX3 SES',
'ZXW SES',
'awd limited',
'FX4 Level II',
'EX Special Edition',
'LX Special Edition PZEV',
'awd ex-l',
'S 7 Passenger',
```

```
'GLS Special Value',
'4.2',
'Luxury 5 Passenger',
'Unlimited',
'4wd s',
'C320 Luxury 4MATIC',
'SP23 Special Edition',
'Trailhawk',
'3.5 SE-R',
'LS Premium',
'GTS V6',
'S500 4MATIC',
'Off-Road',
'SE Off Road',
'Arc 2.3T',
'2.5 GT',
'awd xt',
'awd x w/prem',
'1',
'GLI 1.8T',
'GL PZEV',
'3.0 Avant quattro',
'4x2 6c cxl',
'4.8is',
'Z71 LS',
'Xtreme',
'2500 LS',
'fwd',
'v6 lxi',
'GTC',
'v6',
'awd 2004.5',
'v6 es',
'XLS Sport',
'4x4 v6',
'Summit',
'Edge',
'Tremor',
'ZTS',
'Mach 1 Premium',
'Z71 SLE',
'SVT',
'SVT Cobra',
'ZTW',
'ex v-6',
'S 5-Passenger',
'C240 4MATIC',
```

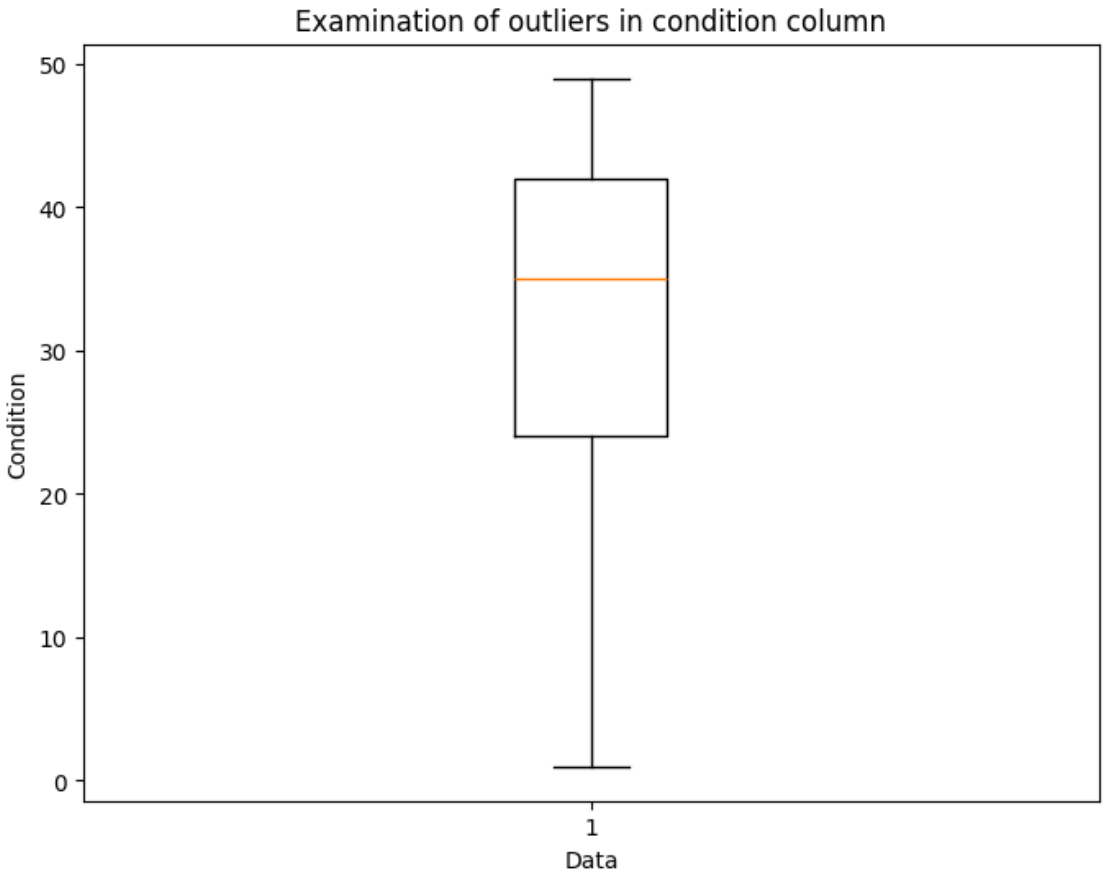
```
 'SL600',
 'LS Ultimate',
 'Sport Edition',
 'GT2',
 'H6-3.0 L.L. Bean Edition',
 'Tiptronic',
 'H6-3.0 VDC',
 'GL 1.8T',
 'GLI VR6',
 'HPT',
 'XRS',
 'R',
 '3.2 Type-S',
 'LX Hybrid',
 'ZR2',
 'v6 lx',
 'LX Value',
 '4x4 v6 xlt pop 2',
 'LX Premium',
 'XLT FX4',
 'SVT Lightning',
 'se7',
 'B2300',
 'C320 4MATIC',
 'C320 Sport',
 'S430 4MATIC',
 'CLK430',
 'ES V6',
 'SSEi',
 'Arc 3.0t',
 'L200',
 'carrera 4x2 carrera 2',
 'L.L. Bean Edition',
 '180hp quattro',
 '2.9',
 '225hp quattro',
 'T6 Elite',
 '4.6is',
 'STS',
 'Z24',
 'eC',
 …]
```

[32]:
```
df['trim'] = df['trim'].apply(lambda x: x.capitalize() if x.islower() else x)
df['trim'] = df['trim'].apply(lambda x: x.upper() if len(x) <= 3 else x)
```

8 Adjusting condition column

[33]:
```
condition_to_plot = df.loc[~df['condition'].isna(),'condition']
```

[34]:
```
plt.figure(figsize=(8, 6))
plt.boxplot(condition_to_plot)

plt.title("Examination of outliers in condition column")
plt.ylabel("Condition")
plt.xlabel("Data")
plt.show()
```



[35]:
```
median_for_model = df.groupby(["make","model"]).agg({"condition":"median"})
median_for_model.reset_index(inplace=True)
median_for_model
```

[35]:

	make	model	condition
0	Acura	CL	19.0
1	Acura	ILX	38.0

```
2     Acura  Integra        2.0
3     Acura   Legend        1.0
4     Acura      MDX       34.0
..      …       …         …
771   Volvo      V70       22.0
772   Volvo       XC       27.0
773   Volvo     XC60       42.0
774   Volvo     XC70       37.5
775   Volvo     XC90       32.0

[776 rows x 3 columns]
```

[36]:
```
df[df['condition'].isna()]
```

[36]:
```
         year        make        model       trim         body transmission  \
14       2014   Chevrolet        Cruze         2LT        Sedan    Automatic
16       2015     Hyundai       Sonata          SE        Sedan    Automatic
22       2014   Chevrolet       Camaro          LT  Convertible    Automatic
25       2015     Hyundai       Sonata          SE        Sedan    Automatic
28       2014         Bmw           X5  sDrive35i          Suv    Automatic
…        …          …           …         …           …           …
43570    2010       Acura          RDX     SH-AWD          Suv    Automatic
43579    2011      Toyota       Tundra     Tundra   Double cab    Automatic
43580    2011  Volkswagen        Jetta    SE PZEV        Sedan    Automatic
182059   2005        Ford        F-150         XL     Supercab    Automatic
541568   2005   Chevrolet  Astro Cargo       Base      Minivan    Automatic

        state  condition  odometer   color interior  \
14         CA        NaN   15686.0    blue    black
16         CA        NaN    8311.0     red        -
22         CA        NaN   33450.0   black    black
25         CA        NaN    9281.0  silver     gray
28         CA        NaN   11278.0    gray    black
…       …         …        …      …       …
43570      OH        NaN   63454.0   black     gray
43579      TX        NaN   46683.0   black    brown
43580      OH        NaN   30778.0   black        -
182059     ON        NaN  264448.0   white     gray
541568     ON        NaN  126798.0     red     gray

                                                  seller      mmr  \
14                                  avis rac/san leandro  13900.0
16                                             avis tra   15200.0
22                                  avis rac/san leandro  20100.0
25      enterprise vehicle exchange / tra / rental / t…  15150.0
28                                  avis rac/san leandro  50400.0
…                                                    …      …
```

```
43570                                    thornhill acura  17800.0
43579                        avalanche preowned vehicles lp  18500.0
43580                                  colonial auto sales   9725.0
182059                                         auto autos   2250.0
541568                                  jim pattison lease   4925.0

        sellingprice                                  saledate
14           10600.0  Tue Dec 16 2014 12:00:00 GMT-0800 (PST)
16            4200.0  Tue Dec 16 2014 13:00:00 GMT-0800 (PST)
22           14700.0  Tue Dec 16 2014 12:00:00 GMT-0800 (PST)
25            8500.0  Tue Dec 16 2014 13:00:00 GMT-0800 (PST)
28           34000.0  Tue Dec 16 2014 13:00:00 GMT-0800 (PST)
…                …                                       …
43570        14500.0  Tue Dec 23 2014 09:30:00 GMT-0800 (PST)
43579        19250.0  Thu Dec 18 2014 12:00:00 GMT-0800 (PST)
43580         9500.0  Tue Dec 23 2014 09:30:00 GMT-0800 (PST)
182059        1000.0  Thu Jan 22 2015 10:30:00 GMT-0800 (PST)
541568        3300.0  Tue Jun 16 2015 03:00:00 GMT-0700 (PDT)

[11286 rows x 15 columns]
```

[37]:
```
merged_df = df.merge(median_for_model, on=["make", "model"], how="left",␣
 ↪suffixes=("", "_median"))

# Fill missing condition values with median values where available
mask = merged_df['condition'].isna()
merged_df.loc[mask, 'condition'] = merged_df.loc[mask, 'condition_median']

# Drop the '_median' column as it is no longer needed
merged_df.drop(columns='condition_median', inplace=True)
merged_df.dropna(subset = ['condition'],inplace = True) #for the rest unmatched␣
 ↪data
df = merged_df
```

[38]:
```
df
```

[38]:
```
        year    make                model         trim       body  \
0       2015     Kia              Sorento           LX        Suv
1       2015     Kia              Sorento           LX        Suv
2       2014     Bmw            3 Series   328i SULEV      Sedan
3       2015   Volvo                 S60           T5      Sedan
4       2014     Bmw  6 Series Gran Coupe         650i      Sedan
…       …      …                    …           …         …
545836  2015     Kia                K900       Luxury      Sedan
545837  2012     Ram                2500  Power Wagon   Crew cab
545838  2012     Bmw                  X5    xDrive35d        Suv
545839  2015  Nissan              Altima        2.5 S      Sedan
```

```
545840  2014    Ford                F-150          XLT  Supercrew

       transmission state  condition  odometer   color interior  \
0         Automatic    CA        5.0   16639.0   white    black
1         Automatic    CA        5.0    9393.0   white    beige
2         Automatic    CA       45.0    1331.0    gray    black
3         Automatic    CA       41.0   14282.0   white    black
4         Automatic    CA       43.0    2641.0    gray    black
…               …   …         …       …      …       …
545836    Automatic    IN       45.0   18255.0  silver    black
545837    Automatic    WA        5.0   54393.0   white    black
545838    Automatic    CA       48.0   50561.0   black    black
545839    Automatic    GA       38.0   16658.0   white    black
545840    Automatic    CA       34.0   15008.0    gray     gray

                                                 seller      mmr  \
0                                kia motors america  inc  20500.0
1                                kia motors america  inc  20800.0
2                  financial services remarketing (lease)  31900.0
3                                volvo na rep/world omni  27500.0
4                  financial services remarketing (lease)  66000.0
…                                                      …      …
545836                                   avis corporation  35300.0
545837                                    i -5 uhlmann rv  30200.0
545838             financial services remarketing (lease)  29800.0
545839  enterprise vehicle exchange / tra / rental / t…  15100.0
545840                   ford motor credit company llc pd  29600.0

        sellingprice                                 saledate
0            21500.0  Tue Dec 16 2014 12:30:00 GMT-0800 (PST)
1            21500.0  Tue Dec 16 2014 12:30:00 GMT-0800 (PST)
2            30000.0  Thu Jan 15 2015 04:30:00 GMT-0800 (PST)
3            27750.0  Thu Jan 29 2015 04:30:00 GMT-0800 (PST)
4            67000.0  Thu Dec 18 2014 12:30:00 GMT-0800 (PST)
…              …                                       …
545836       33000.0  Thu Jul 09 2015 07:00:00 GMT-0700 (PDT)
545837       30800.0  Wed Jul 08 2015 09:30:00 GMT-0700 (PDT)
545838       34000.0  Wed Jul 08 2015 09:30:00 GMT-0700 (PDT)
545839       11100.0  Thu Jul 09 2015 06:45:00 GMT-0700 (PDT)
545840       26700.0  Thu May 28 2015 05:30:00 GMT-0700 (PDT)

[545838 rows x 15 columns]
```

9 Adjusting color and interior

[39]:
```
cols = ['color',"interior"]
for col in cols:
    df[col] = df[col].str.capitalize()
```

Adjusting mmr,sellingprice

[40]:
```
df[df['mmr'].isna()]
```

[40]:

	year	make	model	trim	body	transmission	state	condition	odometer	color	interior	seller	mmr	sellingprice	saledate
299317	2013	Hyundai	Sonata	GLS	Sedan	Automatic	IL	19.0	37254.0	Gray	Gray	kfl llc	NaN	NaN	NaN
396896	2013	Chrysler	300	Base	Sedan	Automatic	IL	27.0	44208.0	Silver	Black	kfl llc	NaN	NaN	NaN
412417	2013	Hyundai	Sonata	GLS	Sedan	Automatic	AZ	28.0	44299.0	Blue	Gray	kfl llc	NaN	NaN	NaN
419592	2013	Chrysler	200	Touring	Sedan	Automatic	IL	25.0	47344.0	White	Black	kfl llc	NaN	NaN	NaN
419608	2013	Dodge	Avenger	SXT	Sedan	Automatic	IL	34.0	44525.0	Gray	Black	kfl llc	NaN	NaN	NaN
419734	2013	Hyundai	Accent	GLS	Sedan	Automatic	IL	3.0	43135.0	Silver	Gray	kfl llc	NaN	NaN	NaN
421822	2013	Kia	Soul	Base	Wagon	Automatic	VA	19.0	29465.0	Silver	Black	kfl llc	NaN	NaN	NaN
446365	2014	Hyundai	Elantra	SE	Sedan	Automatic	IL	19.0	20775.0	White	-	kfl llc	NaN	NaN	NaN
447133	2013	Dodge	Charger	SE	Sedan	Automatic	IL	19.0	45355.0	White	Black	kfl llc	NaN	NaN	NaN
510362	2012	Ford	Fusion	SEL	Sedan	Automatic	VA	26.0	51648.0	Black	Gray	kfl llc	NaN	NaN	NaN
525325	2012	Kia	Forte	LX	Sedan	Automatic	MD	25.0	30927.0	Silver	Gray	kfl llc	NaN	NaN	NaN
530558	2014	Hyundai	Accent	GLS	Sedan	Automatic	IL	19.0	26379.0	Gray	Gray	kfl llc	NaN	NaN	NaN

- Since amount of data is not big,we can drop it

[41]:
```
df.dropna(subset = ['mmr','sellingprice','saledate'],inplace = True)
```

10 Adjusting saledate

[42]:
```
df['saledate'] = pd.to_datetime(df['saledate'],utc=True)
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_20268\165830538.py:1: UserWarning:
Could not infer format, so each element will be parsed individually, falling
```

```
back to `dateutil`. To ensure parsing is consistent and as-expected, please
specify a format.
  df['saledate'] = pd.to_datetime(df['saledate'],utc=True)
```

[43]:
```
df['saledate'] = pd.to_datetime(df['saledate'].dt.date)
```

- Are there any cars,which were manufactured after seladate?
- We consider them as invalid data

[44]:
```
df[df['saledate'].dt.year < df['year']]
```

[44]:

	year	make	model	trim	body	transmission	state	condition	odometer	color	interior	seller	mmr
0	2015	Kia	Sorento	LX	Suv	Automatic	CA	5.0	16639.0	White	Black	kia motors america inc	20500.0
1	2015	Kia	Sorento	LX	Suv	Automatic	CA	5.0	9393.0	White	Beige	kia motors america inc	20800.0
5	2015	Nissan	Altima	2.5 S	Sedan	Automatic	CA	1.0	5554.0	Gray	Black	enterprise vehicle exchange / tra / rental / t…	15350.0
11	2015	Kia	Optima	LX	Sedan	Automatic	CA	48.0	2034.0	Red	Tan	kia motors finance	15150.0
13	2015	Kia	Sorento	LX	Suv	Automatic	CA	5.0	14634.0	Silver	Black	kia motors america inc	20600.0
…	…	…	…	…	…	…	…	…	…	…	…	…	…
67688	2015	Nissan	Altima	2.5 S	Sedan	Automatic	PA	1.0	9399.0	Red	Black	enterprise veh exchange/rental	14950.0
68930	2015	Chevrolet	Tahoe	LTZ	Suv	Automatic	NE	41.0	6018.0	White	Tan	husker auto group inc	55700.0
68932	2015	Chrysler	200	Limited	Sedan	Automatic	IN	2.0	73.0	Gray	Tan	enterprise vehicle exchange / tra / rental / t…	16050.0
71458	2015	Ford	Fusion	SE	Sedan	Automatic	TX	5.0	3427.0	White	Black	rlb investments	13700.0
71459	2015	Hyundai	Sonata	SE	Sedan	Automatic	CA	1.0	9622.0	Black	-		

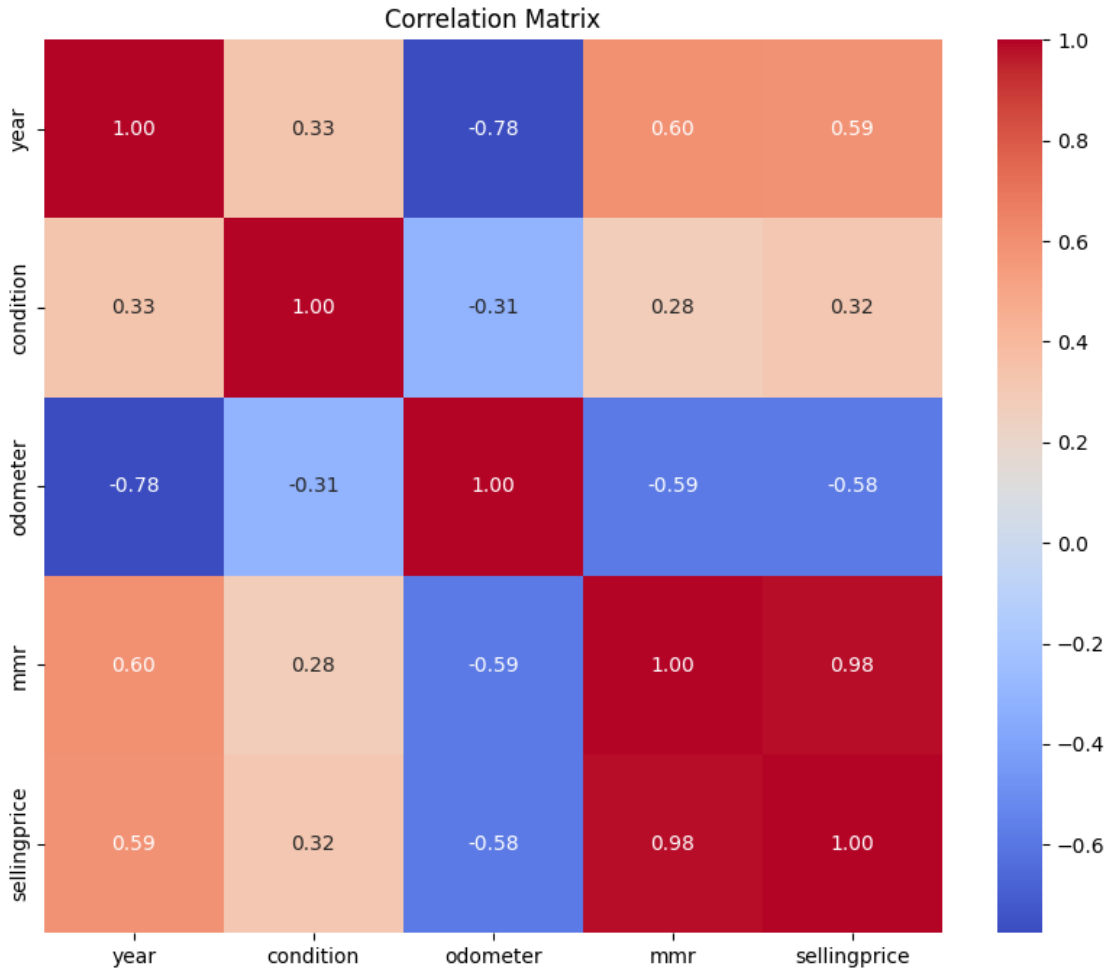
```
71459  enterprise vehicle exchange / tra / rental / t…  15450.0

       sellingprice   saledate
0           21500.0 2014-12-16
1           21500.0 2014-12-16
5           10900.0 2014-12-30
11          17700.0 2014-12-16
13          21500.0 2014-12-16
…               …          …
67688        3600.0 2014-12-30
68930       54600.0 2014-12-31
68932        6600.0 2014-12-31
71458       16300.0 2014-12-31
71459        8900.0 2014-12-31

[201 rows x 15 columns]
```

[45]:
```
df = df[~(df['saledate'].dt.year < df['year'])]
```

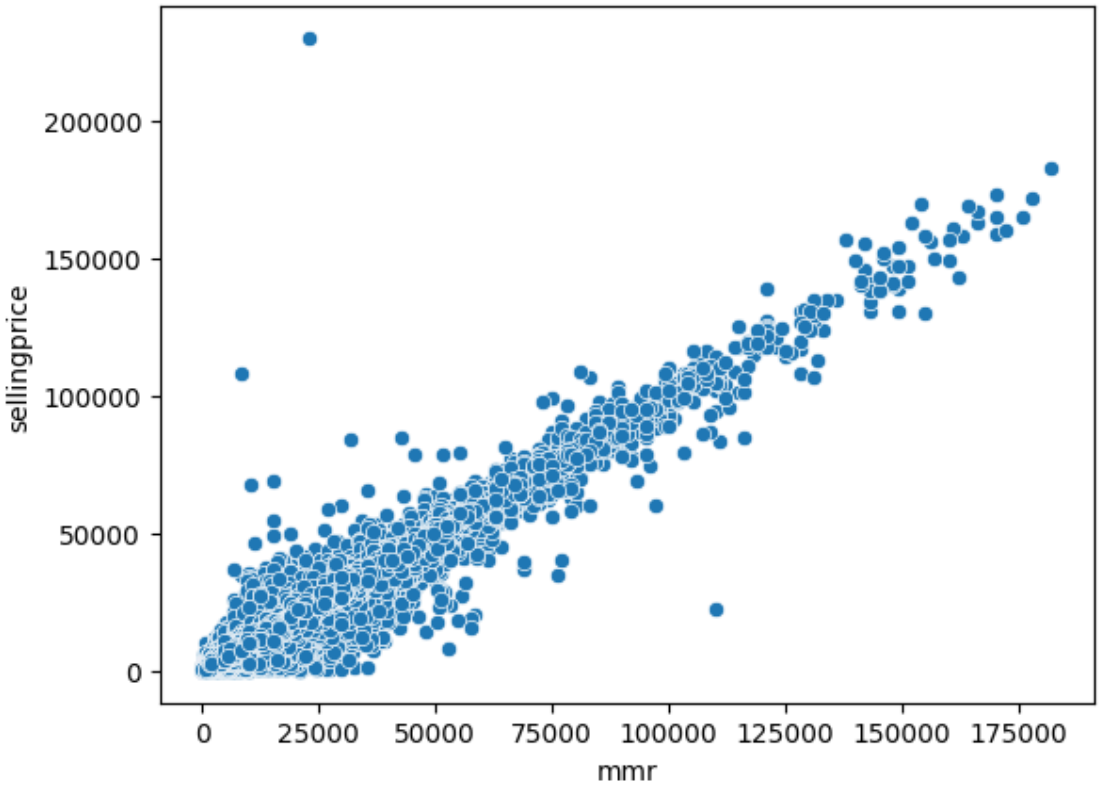
11 Let's see the correlation of factors,which define the sellingprice

[46]:
```
factors = df[["year","condition","odometer","mmr",'sellingprice']]
```

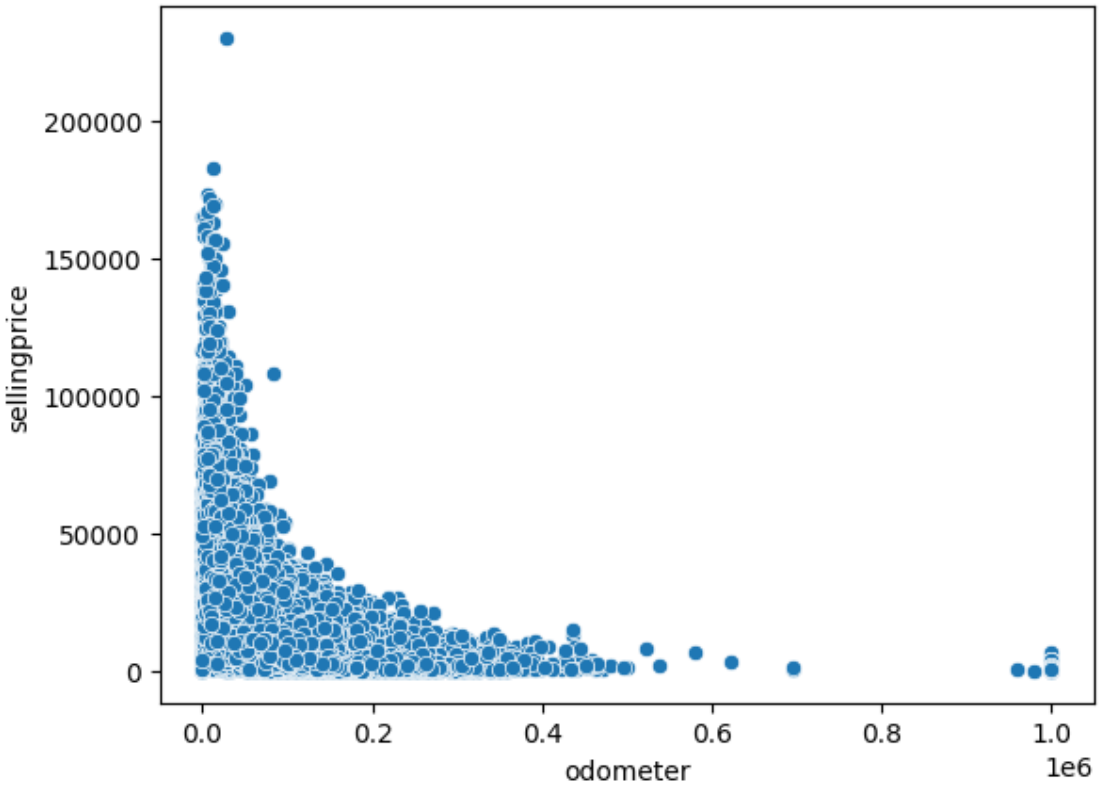
[47]:
```
corr_matrix = factors.corr()

plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()
```



[48]:
```
sns.scatterplot(x='mmr', y='sellingprice', data=df)
plt.show()
```



[49]:
```
sns.scatterplot(x='odometer', y='sellingprice', data=df)
plt.show()
```



- Here scatterplots explain some relationship between sellingprice and mmr and odometer.As you can see,the less number on the odometer is the higher is price,as well Manheim Market Report gives accurate estimation of real price of a car

12 Saving the cleaned df to Excel and analyzing in Tableau

- Link to viz in Tableau - https://public.tableau.com/app/profile/vladislav.zabrovsky/viz/CarMarketAnalysis_

[50]:
```
df.to_excel('cleaned_car_data.xlsx', sheet_name='data', index=False)
```