vehiclesalesdata

April 21, 2025

[44]:
```
import pandas as pd
import numpy as np
```

[45]:
```
car = pd.read_csv("/kaggle/input/vehicle-sales-data/car_prices.csv")
car.sample(5)
```

[45]:

	year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior	seller	mmr	sellingprice	saledate
72228	2006	Jeep	Grand Cherokee	Limited	SUV	automatic	1j4hr58286c361361	mi	49.0	82557.0	beige	tan	kevins marysville auto sales inc	10200.0	11300.0	Thu Jan 08 2015 09:30:00 GMT-0800 (PST)
481530	1992	Buick	LeSabre	Custom	sedan	automatic	1g4hp53l3nh439036	nc	2.0	127745.0	white	-	hendrick kia of cary	450.0	500.0	Mon Jun 01 2015 02:15:00 GMT-0700 (PDT)
205083	2004	Toyota	Corolla	LE	Sedan	automatic	2t1br32e54c215507	sc	19.0	102082.0	gold	beige	east coast honda	4150.0	4100.0	Wed Jan 28 2015 02:00:00 GMT-0800 (PST)
472648	2012	Lincoln	MKS	Base	sedan	automatic	1lnhl9er0cg802853	mi	35.0	47974.0	blue	black	platinum motor cars	19750.0	18900.0	Thu May 28 2015 02:30:00 GMT-0700 (PDT)
129205	2006	Toyota	Corolla	CE	Sedan	NaN	1nxbr32e06z655135	tx	1.0	148606.0	gray	gray	titlebucks/phoenix az8	4125.0	2800.0	Tue Jan 27 2015 03:00:00 GMT-0800 (PST)

[46]:
```
car.dtypes
```

[46]:
```
year              int64
make             object
model            object
trim             object
body             object
transmission     object
vin              object
state            object
condition       float64
odometer        float64
color            object
interior         object
seller           object
mmr             float64
sellingprice    float64
saledate         object
dtype: object
```

[47]:
```
car.shape
```

[47]:
```
(558837, 16)
```

[48]:
```
car.isnull().sum()
```

[48]:
```
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
```

this dataset is too huge so even if we remove 10-12k null value rows, it won't impact it!

[49]:
```
car.dropna(inplace=True)
car.shape
```

[49]: (472325, 16)

[50]:
```
car.duplicated().sum()
```

[50]: 0

There are no duplicate rows in this dataset

[52]:
```
def standardize_text(car, column_name):
    car[column_name] = car[column_name].astype(str).str.lower().str.strip().str.
 ↪replace(r'\s+', ' ', regex=True)
    return car

standardize = ['make', 'model', 'trim', 'body', 'transmission', 'state',␣
 ↪'seller']

for col in standardize:
    car = standardize_text(car, col)
car.head()
```

[52]:

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

here we standardized all text columns, we made all strings lowercase, removed all whitespace from beginning & end of text & also replaced extra whitespaces with one single space.

[53]:
```
car['saledate'] = car['saledate'].astype(str)
car['saledate'] = car['saledate'].str.slice(4, 15)
car.head()
```

[53]:

	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	Dec 16 2014
1	2015	kia	sorento	lx	suv	automatic	5xyktca69fg561319	ca	5.0	9393.0	white	beige	kia motors america inc	20800.0	21500.0	Dec 16 2014
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	Jan 15 2015
3	2015	volvo	s60	t5	sedan	automatic	yv1612tb4f1310987	ca	41.0	14282.0	white	black	volvo na rep/world omni	27500.0	27750.0	Jan 29 2015
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	Dec 18 2014

first we extract the exact date in saledate columns then we'll be applying further functions.

[54]:
```
car['saledate'] = pd.to_datetime(car['saledate'], format='%b %d %Y',␣
 ↪errors='coerce')
car['saledate'] = car['saledate'].dt.strftime('%d-%m-%Y')
car.head()
```

[54]:

	year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior
0	2015	kia	sorento	lx	suv	automatic	5xyktca69fg566472	ca	5.0	16639.0	white	black
1	2015	kia	sorento	lx	suv	automatic	5xyktca69fg561319	ca	5.0	9393.0	white	beige
2	2014	bmw	3 series	328i sulev	sedan	automatic	wba3c1c51ek116351	ca	45.0	1331.0	gray	black
3	2015	volvo	s60	t5	sedan	automatic						
4	2014	bmw	6 series gran coupe	650i	sedan	automatic						

```
3  yv1612tb4f1310987    ca       41.0   14282.0  white    black
4  wba6b2c57ed129731    ca       43.0    2641.0   gray    black

                                   seller      mmr  sellingprice    saledate
0                  kia motors america inc  20500.0       21500.0  16-12-2014
1                  kia motors america inc  20800.0       21500.0  16-12-2014
2  financial services remarketing (lease)  31900.0       30000.0  15-01-2015
3                volvo na rep/world omni   27500.0       27750.0  29-01-2015
4  financial services remarketing (lease)  66000.0       67000.0  18-12-2014
```

here we can see saledate column in proper dd-mm-yyyy format

[56]:
```
car.columns = car.columns.str.strip().str.lower().str.replace(' ', '_')
car.head()
```

[56]:
```
   year   make                model        trim   body transmission  \
0  2015    kia              sorento          lx    suv    automatic
1  2015    kia              sorento          lx    suv    automatic
2  2014    bmw            3 series  328i sulev  sedan    automatic
3  2015  volvo                 s60          t5  sedan    automatic
4  2014    bmw  6 series gran coupe        650i  sedan    automatic

                 vin state  condition  odometer  color interior  \
0  5xyktca69fg566472    ca        5.0   16639.0  white    black
1  5xyktca69fg561319    ca        5.0    9393.0  white    beige
2  wba3c1c51ek116351    ca       45.0    1331.0   gray    black
3  yv1612tb4f1310987    ca       41.0   14282.0  white    black
4  wba6b2c57ed129731    ca       43.0    2641.0   gray    black

                                   seller      mmr  sellingprice    saledate
0                  kia motors america inc  20500.0       21500.0  16-12-2014
1                  kia motors america inc  20800.0       21500.0  16-12-2014
2  financial services remarketing (lease)  31900.0       30000.0  15-01-2015
3                volvo na rep/world omni   27500.0       27750.0  29-01-2015
4  financial services remarketing (lease)  66000.0       67000.0  18-12-2014
```

here we renamed column header to be clean and uniform

[59]:
```
car['saledate'] = pd.to_datetime(car['saledate'])
car.dtypes
```

```
/tmp/ipykernel_31/674071443.py:1: UserWarning: Parsing dates in %d-%m-%Y format
when dayfirst=False (the default) was specified. Pass `dayfirst=True` or specify
a format to silence this warning.
  car['saledate'] = pd.to_datetime(car['saledate'])
```

[59]:
```
year                  int64
make                 object
```

```
model                    object
trim                     object
body                     object
transmission             object
vin                      object
state                    object
condition               float64
odometer                float64
color                    object
interior                 object
seller                   object
mmr                     float64
sellingprice            float64
saledate        datetime64[ns]
dtype: object
```

at last we again checked if all columns have fix data types