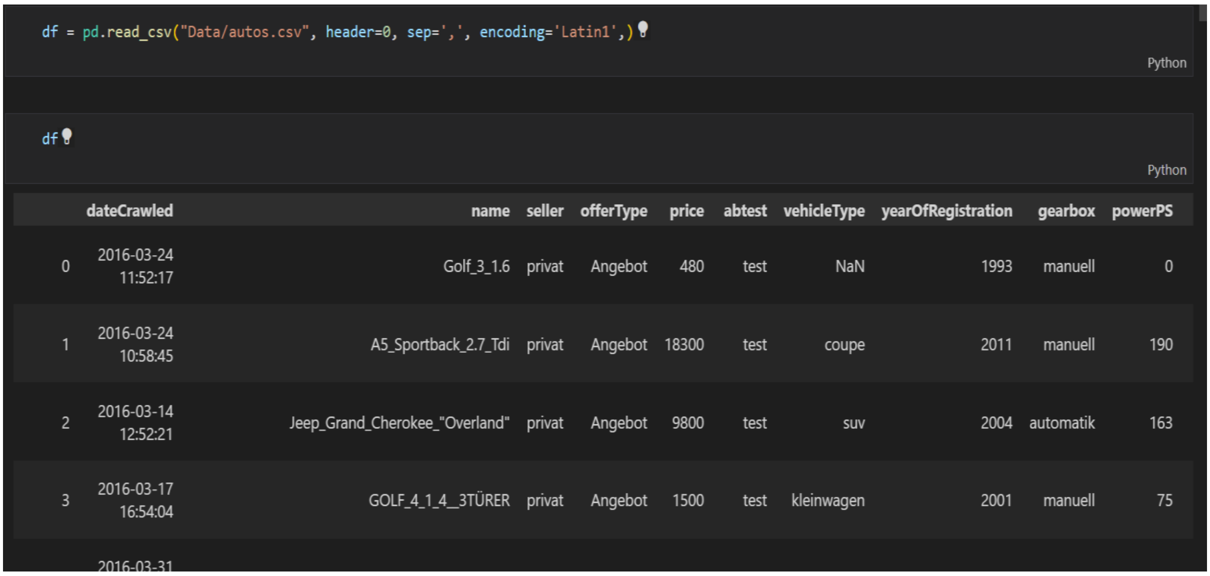
load data set

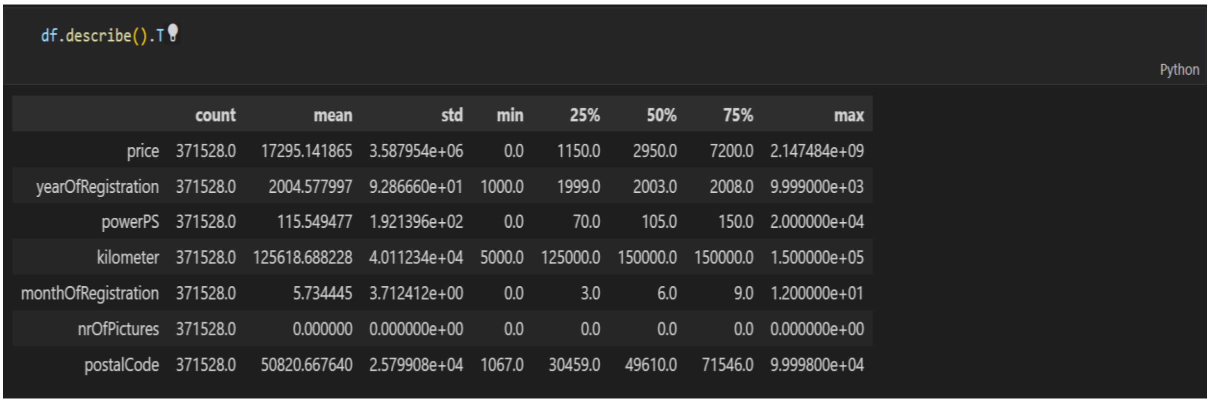


```
df = pd.read_csv("Data/autos.csv", header=0, sep=',', encoding='Latin1',)
```

```
df
```

	dateCrawled	name	seller	offerType	price	abtest	vehicleType	yearOfRegistration	gearbox	powerPS
0	2016-03-24 11:52:17	Golf_3_1.6	privat	Angebot	480	test	NaN	1993	manuell	0
1	2016-03-24 10:58:45	A5_Sportback_2.7_Tdi	privat	Angebot	18300	test	coupe	2011	manuell	190
2	2016-03-14 12:52:21	Jeep_Grand_Cherokee_"Overland"	privat	Angebot	9800	test	suv	2004	automatik	163
3	2016-03-17 16:54:04	GOLF_4_1_4__3TÜRER	privat	Angebot	1500	test	kleinwagen	2001	manuell	75
	2016-03-31									

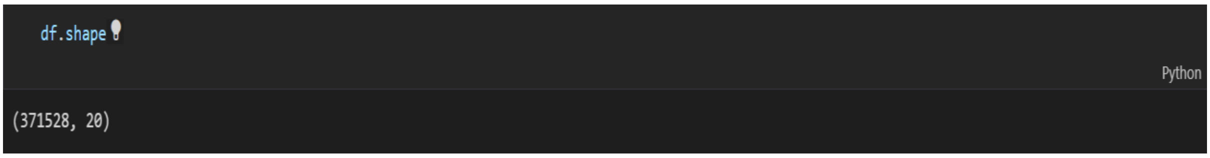
Describe the datas



```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
price	371528.0	17295.141865	3.587954e+06	0.0	1150.0	2950.0	7200.0	2.147484e+09
yearOfRegistration	371528.0	2004.577997	9.286660e+01	1000.0	1999.0	2003.0	2008.0	9.999000e+03
powerPS	371528.0	115.549477	1.921396e+02	0.0	70.0	105.0	150.0	2.000000e+04
kilometer	371528.0	125618.688228	4.011234e+04	5000.0	125000.0	150000.0	150000.0	1.500000e+05
monthOfRegistration	371528.0	5.734445	3.712412e+00	0.0	3.0	6.0	9.0	1.200000e+01
nrOfPictures	371528.0	0.000000	0.000000e+00	0.0	0.0	0.0	0.0	0.000000e+00
postalCode	371528.0	50820.667640	2.579908e+04	1067.0	30459.0	49610.0	71546.0	9.999800e+04

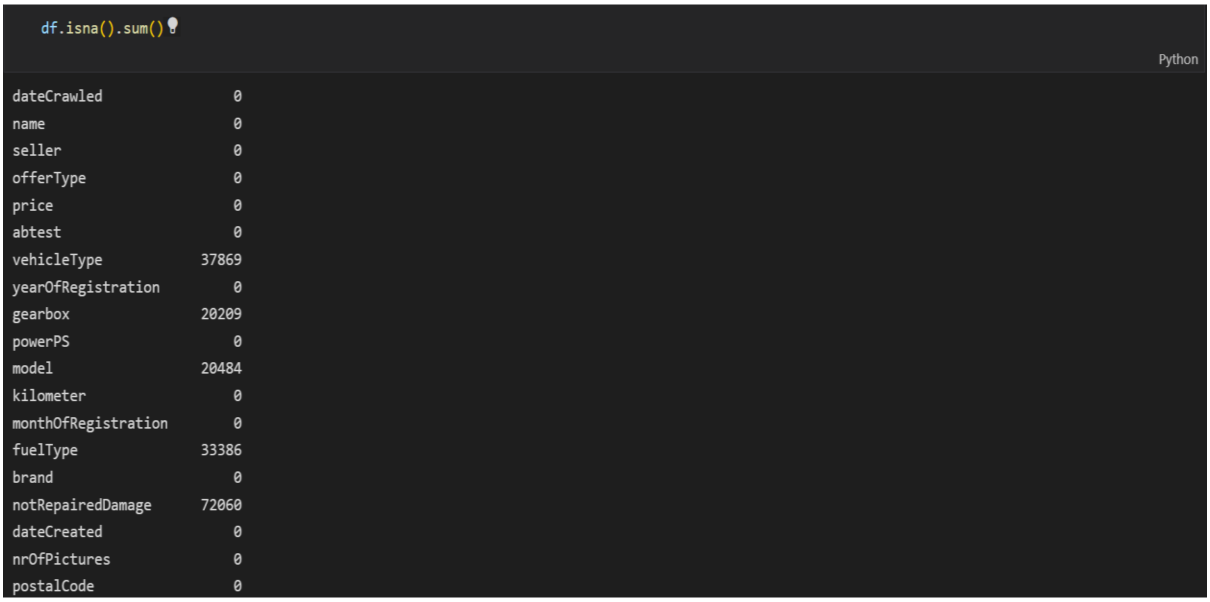
Shape of dataset



```
df.shape
```

```
(371528, 20)
```

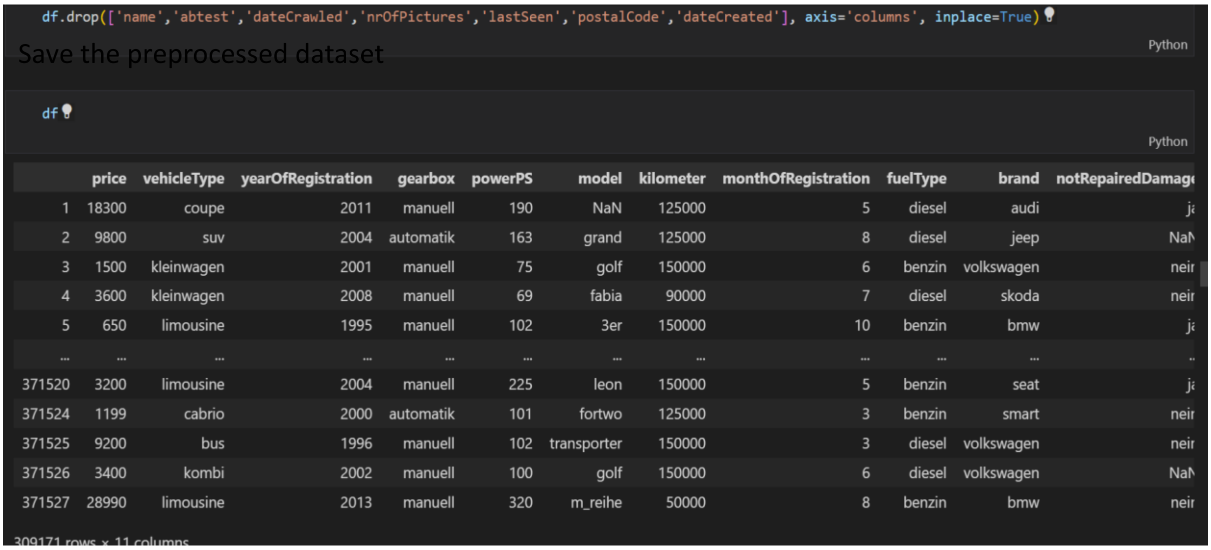
Find null values



```
df.isna().sum()
```

```
dateCrawled                  0
name                         0
seller                       0
offerType                    0
price                        0
abtest                       0
vehicleType              37869
yearOfRegistration           0
gearbox                  20209
powerPS                      0
model                    20484
kilometer                    0
monthOfRegistration          0
fuelType                 33386
brand                        0
notRepairedDamage        72060
dateCreated                  0
nrOfPictures                 0
postalCode                   0
```

Drop some datas



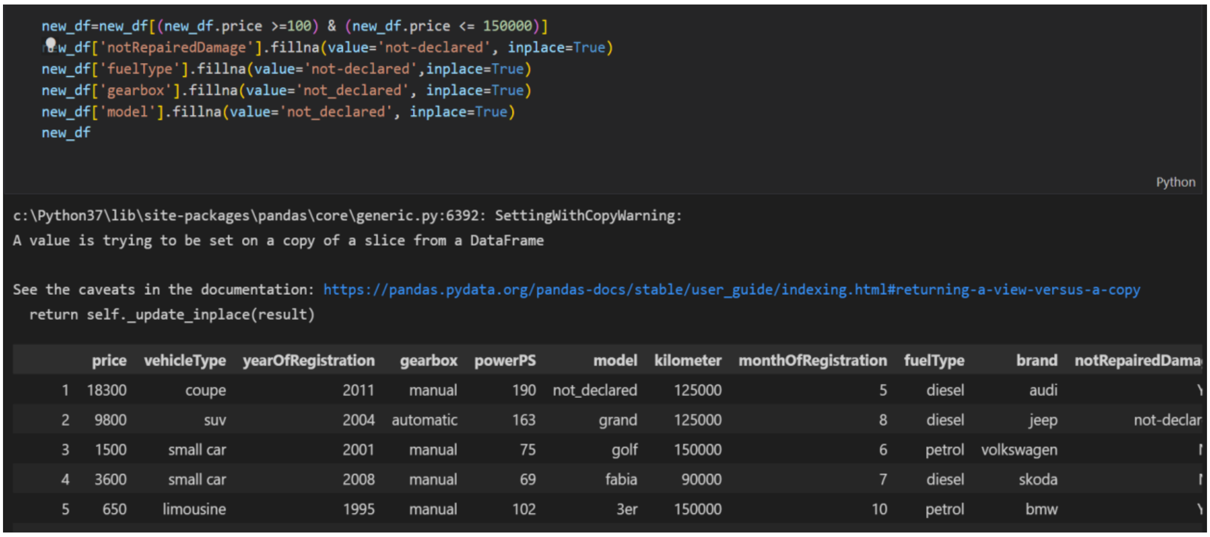
```
df.drop(['name','abtest','dateCrawled','nrOfPictures','lastSeen','postalCode','dateCreated'], axis='columns', inplace=True)
```

Save the preprocessed dataset

```
df
```

	price	vehicleType	yearOfRegistration	gearbox	powerPS	model	kilometer	monthOfRegistration	fuelType	brand	notRepairedDamage
1	18300	coupe	2011	manuell	190	NaN	125000	5	diesel	audi	ja
2	9800	suv	2004	automatik	163	grand	125000	8	diesel	jeep	NaN
3	1500	kleinwagen	2001	manuell	75	golf	150000	6	benzin	volkswagen	nein
4	3600	kleinwagen	2008	manuell	69	fabia	90000	7	diesel	skoda	nein
5	650	limousine	1995	manuell	102	3er	150000	10	benzin	bmw	ja
...	...	...	...	...	...	...	...	...	...	...	...
371520	3200	limousine	2004	manuell	225	leon	150000	5	benzin	seat	ja
371524	1199	cabrio	2000	automatik	101	fortwo	125000	3	benzin	smart	nein
371525	9200	bus	1996	manuell	102	transporter	150000	3	diesel	volkswagen	nein
371526	3400	kombi	2002	manuell	100	golf	150000	6	diesel	volkswagen	NaN
371527	28990	limousine	2013	manuell	320	m_reihe	50000	8	benzin	bmw	nein

309171 rows × 11 columns

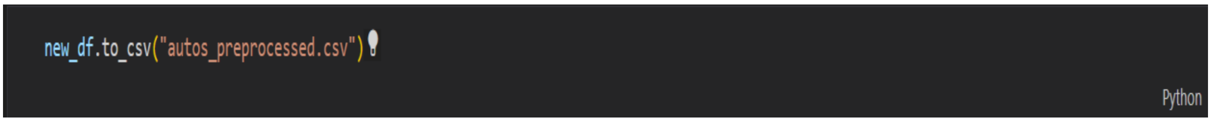


```
new_df=new_df[(new_df.price >=100) & (new_df.price <= 150000)]
new_df['notRepairedDamage'].fillna(value='not-declared', inplace=True)
new_df['fuelType'].fillna(value='not-declared',inplace=True)
new_df['gearbox'].fillna(value='not_declared', inplace=True)
new_df['model'].fillna(value='not_declared', inplace=True)
new_df
```

```
c:\Python37\lib\site-packages\pandas\core\generic.py:6392: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  return self._update_inplace(result)
```

	price	vehicleType	yearOfRegistration	gearbox	powerPS	model	kilometer	monthOfRegistration	fuelType	brand	notRepairedDama
1	18300	coupe	2011	manual	190	not_declared	125000	5	diesel	audi	Y
2	9800	suv	2004	automatic	163	grand	125000	8	diesel	jeep	not-declar
3	1500	small car	2001	manual	75	golf	150000	6	petrol	volkswagen	N
4	3600	small car	2008	manual	69	fabia	90000	7	diesel	skoda	N
5	650	limousine	1995	manual	102	3er	150000	10	petrol	bmw	Y



```
new_df.to_csv("autos_preprocessed.csv")
```