Data Preparation

Miguel Alejandro Salas Reyna (2022), Data Science and Mathematics Student.

Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM).

In [1]:
```
# libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import date
import numpy as np
```

In [2]:
```
df = pd.read_csv("Automotive_2.csv")
```

In [3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 171528 entries, 0 to 171527
Data columns (total 20 columns):
 #   Column               Non-Null Count   Dtype 
---  ------               --------------   ----- 
 0   dateCrawled          171528 non-null  object
 1   name                 171528 non-null  object
 2   seller               171528 non-null  object
 3   offerType            171528 non-null  object
 4   price                171528 non-null  object
 5   abtest               171516 non-null  object
 6   vehicleType          154049 non-null  object
 7   yearOfRegistration   171518 non-null  object
 8   gearbox              162178 non-null  object
 9   powerPS              171524 non-null  object
 10  model                162165 non-null  object
 11  kilometer            171528 non-null  int64 
 12  monthOfRegistration  171512 non-null  object
 13  fuelType             156241 non-null  object
 14  brand                171486 non-null  object
 15  notRepairedDamage    138361 non-null  object
 16  dateCreated          171528 non-null  object
 17  nrOfPictures         171528 non-null  int64 
 18  postalCode           171528 non-null  object
 19  lastSeen             171396 non-null  object
dtypes: int64(2), object(18)
memory usage: 26.2+ MB
```

In [4]:
```
df.head(10)
```

Out[4]:

	dateCrawled	name	seller	offerType	price	abt
0	3/26/2016 14:57	BMW_320d_DPFXenon_Tempomat_Sitzheizung_PDC_Kli...	privat	Angebot	10499	con
1	3/23/2016 20:57	Renault_clio_mit_nagelneuen_T?!*	privat	Angebot	1199	t
2	3/19/2016 18:56	VW_LUPO._1.1l_T?_3/18	privat	Angebot	2750	t
3	3/30/2016 12:51	Opel_Astra_Caravan_Sport_sehr_gepflegt	privat	Angebot	3500	con
4	3/26/2016 21:36	Ford_Focus_Turnier_1.6_TDCi_DPF_Trend	privat	Angebot	7500	con
5	3/15/2016 19:45	Audi_Q5_3.0_TDI_quattro_S_tronic	privat	Angebot	34800	con
6	3/7/2016 22:48	Alfa_Romeo_147	privat	Angebot	1050	t
7	3/30/2016 20:57	renaut_espace_fast_frei	privat	Angebot	200	con
8	4/2/2016 10:56	BMW_e36_316i_Compact_1_9_T?_neu	privat	Angebot	1250	t
9	4/1/2016 12:53	Renault_Trafic_2.5_dCi_Generation_Expression	privat	Angebot	5200	t

In [5]:
```
df.shape
```

Out[5]:
```
(171528, 20)
```

In [6]:
```
df['notRepairedDamage'] = df['notRepairedDamage'].fillna(0)
```

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 171528 entries, 0 to 171527
Data columns (total 20 columns):
 #   Column               Non-Null Count   Dtype 
---  ------               --------------   ----- 
 0   dateCrawled          171528 non-null  object
 1   name                 171528 non-null  object
 2   seller               171528 non-null  object
 3   offerType            171528 non-null  object
 4   price                171528 non-null  object
 5   abtest               171516 non-null  object
 6   vehicleType          154049 non-null  object
 7   yearOfRegistration   171518 non-null  object
 8   gearbox              162178 non-null  object
 9   powerPS              171524 non-null  object
 10  model                162165 non-null  object
 11  kilometer            171528 non-null  int64 
 12  monthOfRegistration  171512 non-null  object
 13  fuelType             156241 non-null  object
 14  brand                171486 non-null  object
 15  notRepairedDamage    171528 non-null  object
 16  dateCreated          171528 non-null  object
 17  nrOfPictures         171528 non-null  int64 
 18  postalCode           171528 non-null  object
 19  lastSeen             171396 non-null  object
dtypes: int64(2), object(18)
memory usage: 26.2+ MB
```

In [8]:
```
null=df[df['lastSeen'].isnull()]
null
```

Out[8]:

lel	kilometer	monthOfRegistration	fuelType	brand	notRepairedDamage	dateCreated	nrOfPict
00	5	benzin	audi	no	3/22/2016 0:00	0	8
00	10	benzin	bmw	yes	3/22/2016 0:00	0	45
00	12	diesel	ford	NaN	3/9/2016 0:00	0	53
00	2	benzin	mini	no	3/25/2016 0:00	0	86
00	0	benzin	bmw	NaN	3/24/2016 0:00	0	36
...	...	...	...	...	...	...	
00	2	benzin	mercedes_benz	no	3/20/2016 0:00	0	30
00	6	benzin	fiat	NaN	3/30/2016 0:00	0	76
00	2	NaN	peugeot	yes	4/5/2016 0:00	0	47
00	6	benzin	mini	no	3/8/2016 0:00	0	44
00	0	NaN	bmw	NaN	3/31/2016 0:00	0	22

In [9]:
```
#Drop of null values
df=df.dropna()
```

In [10]:
```
df
```

Out[10]:

	dateCrawled	name	seller	offerType	price
0	3/26/2016 14:57	BMW_320d_DPFXenon_Tempomat_Sitzheizung_PDC_Kli...	privat	Angebot	10499
1	3/23/2016 20:57	Renault_clio_mit_nagelneuen_T?!*	privat	Angebot	1199
2	3/19/2016 18:56	VW_LUPO._1.1l_T?_3/18	privat	Angebot	2750
3	3/30/2016 12:51	Opel_Astra_Caravan_Sport_sehr_gepflegt	privat	Angebot	3500
4	3/26/2016 21:36	Ford_Focus_Turnier_1.6_TDCi_DPF_Trend	privat	Angebot	7500
...	...	...	...	...	...
171521	3/27/2016 20:36	Opel_Zafira_1.6_Elegance_T?_12/16	privat	Angebot	1150
171524	3/5/2016 19:56	Smart_smart_leistungssteigerung_100ps	privat	Angebot	1199
171525	3/19/2016 18:57	Volkswagen_Multivan_T4_TDI_7DC_UY2	privat	Angebot	9200
171526	3/20/2016 19:41	VW_Golf_Kombi_1_9l_TDI	privat	Angebot	3400
171527	3/7/2016 19:39	BMW_M135i_vollausgestattet_NP_52.720____Euro	privat	Angebot	28990

138344 rows × 20 columns

In [11]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 138344 entries, 0 to 171527
Data columns (total 20 columns):
 #   Column               Non-Null Count   Dtype 
---  ------               --------------   ----- 
 0   dateCrawled          138344 non-null  object
 1   name                 138344 non-null  object
 2   seller               138344 non-null  object
 3   offerType            138344 non-null  object
 4   price                138344 non-null  object
 5   abtest               138344 non-null  object
 6   vehicleType          138344 non-null  object
 7   yearOfRegistration   138344 non-null  object
 8   gearbox              138344 non-null  object
 9   powerPS              138344 non-null  object
 10  model                138344 non-null  object
 11  kilometer            138344 non-null  int64 
 12  monthOfRegistration  138344 non-null  object
 13  fuelType             138344 non-null  object
 14  brand                138344 non-null  object
 15  notRepairedDamage    138344 non-null  object
 16  dateCreated          138344 non-null  object
 17  nrOfPictures         138344 non-null  int64 
 18  postalCode           138344 non-null  object
 19  lastSeen             138344 non-null  object
dtypes: int64(2), object(18)
memory usage: 22.2+ MB
```

In [12]:
```
df['notRepairedDamage'] = df['notRepairedDamage'].replace("no", 0)
```

In [13]:
```
df['notRepairedDamage'] = df['notRepairedDamage'].replace("yes", 1)
```

In [14]:
```
df["notRepairedDamage"].unique()
```

Out[14]:
```
array([0, 1], dtype=int64)
```

In [15]:
```
df["seller"].unique()
```

Out[15]:
```
array(['privat'], dtype=object)
```

In [16]:
```
df["offerType"].unique()
```

Out[16]:
```
array(['Angebot', 'Gesuch'], dtype=object)
```

In [17]:
```
columns = df.columns
list(columns)
```

Out[17]:
```
['dateCrawled',
 'name',
 'seller',
 'offerType',
 'price',
 'abtest',
 'vehicleType',
 'yearOfRegistration',
 'gearbox',
 'powerPS',
 'model',
 'kilometer',
 'monthOfRegistration',
 'fuelType',
 'brand',
 'notRepairedDamage',
 'dateCreated',
 'nrOfPictures',
 'postalCode',
 'lastSeen']
```

In [18]:
```
for col in df:
    print(col + " ")
    print(df[col].unique())
```

```
'47' '284' '273' '34' '258' '158' '233' '309' '307' '370' '435' '260'
'172' '408' '196' '219' '62' '43' '315' '1595' '274' '296' '455' '377'
'285' '776' '238' '476' '367' '360' '27' '1400' '445' '316' '17700' '432'
'217' '202' '350' '290' '352' '49' '1' '349' '271' '148' '298' '322' '2'
'186' '11011' '672' '325' '351' '127' '507' '571' '142' '213' '46' '201'
'371' '161' '203' '407' '345' '93' '405' '421' '1199' '487' '236' '382'
'295' '374' '388' '42' '521' '310' '1598' '525' '292' '328' '11' '149'
'380' '379' '223' '183' '23' '357' '279' '1162' '162' '276' '266' '157'
'386' '222' '89' '198' '159' '560' '10522' '327' '15033' '268' '550'
'1896' '401' '254' '603' '1399' '347' '329' '514' '339' '510' '164' '283'
'1000' '247' '24' '6512' '19' '454' '314' '321' '551' '212' '585' '1300'
'25' '604' '999' '381' '168' '6' '123' '243' '348' '14' '289' '181' '253'
'16312' '214' '249' '443' '311' '18' '336' '20' '416' '29' '208' '20000'
'134' '22' '678' '199' '515' '950' '1021' '1500' '138' '1870' '5' '1401'
'303' '751' '242' '237' '11620' '269' '600' '519' '13636' '335' '262'
'15' '398' '30' '396' '9710' '16' '234' '246' '2009' '426' '430' '35'
'6062' '390' '449' '1275' '517' '375' '431' '540' '436' '572' '187' '415'
'481' '244' '362' '399' '17' '1016' '10' '1221' '261' '255' '2402' '555'
'1362' '359' '544' '645' '907' '227' '625' '287' '318' '31' '216' '702'
'1995' '1003' '1600' '1432' '376' '9013' '557' '424' '15017' '485' '2018'
```

In [19]:
```
df['dateCrawled'] =  pd.to_datetime(df['dateCrawled'])
df['dateCreated'] =  pd.to_datetime(df['dateCreated'])
df['lastSeen'] =  pd.to_datetime(df['lastSeen'])
```

In [20]:
```
df['days'] = (df['lastSeen'] - df['dateCrawled']).dt.days
```

In [21]:
```
df = df.drop_duplicates()
```

In [22]:
```
df.describe()
```

Out[22]:

	kilometer	notRepairedDamage	nrOfPictures	days
count	138330.000000	138330.000000	138330.0	138330.000000
mean	125400.093978	0.093949	0.0	8.347618
std	39290.017236	0.291759	0.0	8.334088
min	5000.000000	0.000000	0.0	0.000000
25%	100000.000000	0.000000	0.0	2.000000
50%	150000.000000	0.000000	0.0	5.000000
75%	150000.000000	0.000000	0.0	13.000000
max	150000.000000	1.000000	0.0	33.000000

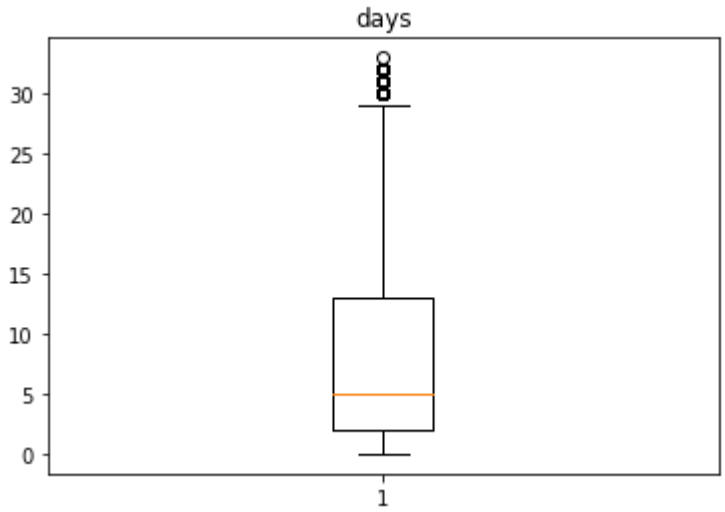
In [23]:
```
df
```

Out[23]:

	dateCrawled	name	seller	offerType	p
0	2016-03-26 14:57:00	BMW_320d_DPFXenon_Tempomat_Sitzheizung_PDC_Kli...	privat	Angebot	10
1	2016-03-23 20:57:00	Renault_clio_mit_nagelneuen_T?!*	privat	Angebot	1
2	2016-03-19 18:56:00	VW_LUPO._1.1l_T?_3/18	privat	Angebot	2
3	2016-03-30 12:51:00	Opel_Astra_Caravan_Sport_sehr_gepflegt	privat	Angebot	3
4	2016-03-26 21:36:00	Ford_Focus_Turnier_1.6_TDCi_DPF_Trend	privat	Angebot	7
...	...	...	...	...	
171521	2016-03-27 20:36:00	Opel_Zafira_1.6_Elegance_T?_12/16	privat	Angebot	1
171524	2016-03-05 19:56:00	Smart_smart_leistungssteigerung_100ps	privat	Angebot	1
171525	2016-03-19 18:57:00	Volkswagen_Multivan_T4_TDI_7DC_UY2	privat	Angebot	9
171526	2016-03-20 19:41:00	VW_Golf_Kombi_1_9l_TDI	privat	Angebot	3
171527	2016-03-07 19:39:00	BMW_M135i_vollausgestattet_NP_52.720____Euro	privat	Angebot	28

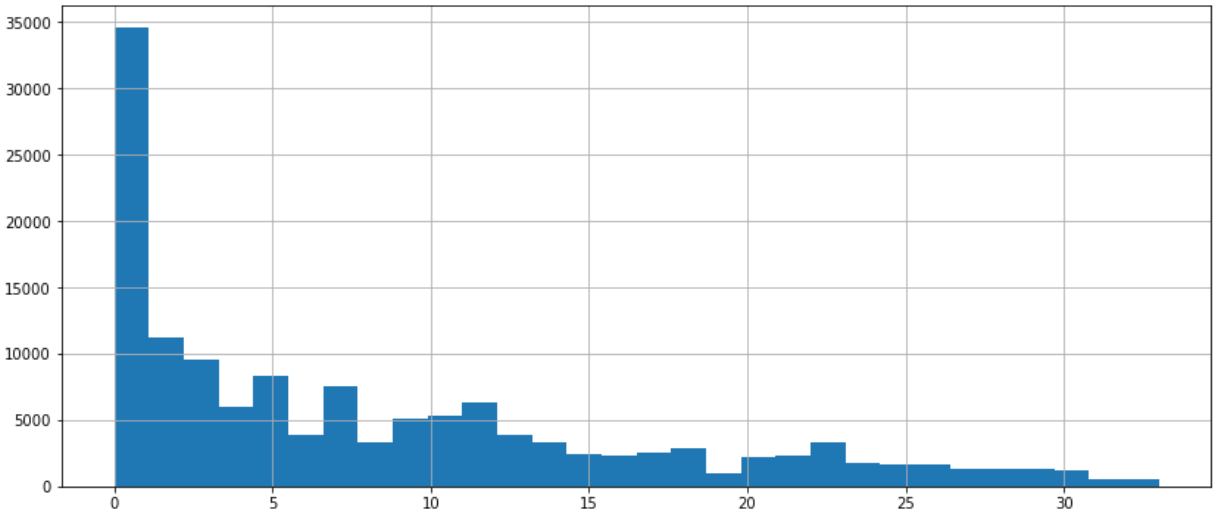
138330 rows × 21 columns

In [24]:
```
plt.title('days')
plt.boxplot(df.days)
plt.show()
```



In [25]:
```
df['days'].hist(bins=30, figsize=[14,6])
```

Out[25]: <AxesSubplot:>



In [26]:
```
df['status'] = np.where(df['days']== 0, 1, 0)
```

```
C:\Users\miigu\AppData\Local\Temp/ipykernel_3788/1346384139.py:1: SettingWithCo
pyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/sta
ble/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pyd
ata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-c
opy)
  df['status'] = np.where(df['days']== 0, 1, 0)
```

In [27]:
```
df.loc[df['days'] == 0]
```

Out[27]:

	dateCrawled	name	seller	offerType	pri
12	2016-03-19 20:51:00	Toyota_Yaris_Verso_1.3_Sol	privat	Angebot	15
17	2016-03-30 00:46:00	Volkswagen_Polo_1.2_Comfortline	privat	Angebot	25
22	2016-03-12 10:55:00	Passat_2.0_TDI_140_PS_140ooo_km_Tempomat_Sitzh...	privat	Angebot	99
41	2016-03-26 10:48:00	Polo_United_1.4_TDI	privat	Angebot	39
49	2016-04-04 16:50:00	Zafira_Tourer_2.0_CDTI_Start_Stop_Navi_Kamera	privat	Angebot	60
...	...	...	...	...	
171454	2016-03-12 17:59:00	Audi_A_6_S_line_.kombi__6_Gang.2_5__Diesel_.__...	privat	Angebot	58
171465	2016-03-10 11:51:00	Vito_108_CDI	privat	Angebot	16
171466	2016-03-08 20:43:00	GOLF_V_1.9_TDI_GT_SPORT_BI_XENON__NAVI_DVD_ROT	privat	Angebot	30
171471	2016-03-07 19:58:00	Trabant_P70_Coupe_Oldtimer_IFA_DDR	privat	Angebot	33
171520	2016-03-19 19:53:00	turbo_defekt	privat	Angebot	32

23568 rows × 22 columns

In [28]:
```
df = df.drop(columns=['seller', 'nrOfPictures'])
```

In [29]:
```
df["price"] = pd.to_numeric(df["price"])
df["yearOfRegistration"] = pd.to_numeric(df["yearOfRegistration"])
df["powerPS"] = pd.to_numeric(df["powerPS"])
df["kilometer"] = pd.to_numeric(df["kilometer"])
df["monthOfRegistration"] = pd.to_numeric(df["monthOfRegistration"])
```

In [30]:
```
df.describe()
```

Out[30]:

	price	yearOfRegistration	powerPS	kilometer	monthOfRegistration	notRe
count	1.383300e+05	138330.000000	138330.000000	138330.000000	138330.000000	
mean	8.083142e+03	2002.926863	123.526039	125400.093978	6.082419	
std	3.843694e+05	6.550282	173.810988	39290.017236	3.536244	
min	0.000000e+00	1910.000000	0.000000	5000.000000	0.000000	
25%	1.399000e+03	1999.000000	75.000000	100000.000000	3.000000	
50%	3.499000e+03	2003.000000	115.000000	150000.000000	6.000000	
75%	7.999000e+03	2007.000000	150.000000	150000.000000	9.000000	
max	1.000000e+08	2018.000000	20000.000000	150000.000000	12.000000	

In [31]:
```
print(df.groupby(['model']).count())
```

```
         dateCrawled  name  offerType  price  abtest  vehicleType  \
model
100              178   178        178    178     178          178
145               16    16         16     16      16           16
147              225   225        225    225     225          225
156              239   239        239    239     239          239
159               91    91         91     91      91           91
...              ...   ...        ...    ...     ...          ...
yaris            436   436        436    436     436          436
yeti              96    96         96     96      96           96
ypsilon           79    79         79     79      79           79
z_reihe          385   385        385    385     385          385
zafira          1153  1153       1153   1153    1153         1153

         yearOfRegistration  gearbox  powerPS  kilometer  monthOfRegistration
\
model
100                     178      178      178        178                  178
145                      16       16       16         16                   16
147                     225      225      225        225                  225
156                     239      239      239        239                  239
159                      91       91       91         91                   91
...                     ...      ...      ...        ...                  ...
yaris                   436      436      436        436                  436
yeti                     96       96       96         96                   96
ypsilon                  79       79       79         79                   79
z_reihe                 385      385      385        385                  385
zafira                 1153     1153     1153       1153                 1153

         fuelType  brand  notRepairedDamage  dateCreated  postalCode  \
model
100           178    178                178          178         178
145            16     16                 16           16          16
147           225    225                225          225         225
156           239    239                239          239         239
159            91     91                 91           91          91
...           ...    ...                ...          ...         ...
yaris         436    436                436          436         436
yeti           96     96                 96           96          96
ypsilon        79     79                 79           79          79
z_reihe       385    385                385          385         385
zafira       1153   1153               1153         1153        1153

         lastSeen  days  status
model
100           178   178     178
145            16    16      16
147           225   225     225
156           239   239     239
159            91    91      91
...           ...   ...     ...
yaris         436   436     436
yeti           96    96      96
ypsilon        79    79      79
```

```
z_reihe       385   385     385
zafira       1153  1153    1153

[250 rows x 19 columns]
```

In [32]:
```
car_models = df.model.value_counts()
```

In [33]:
```
car_models = pd.DataFrame(car_models)
```

In [34]:
```
car_models
```

Out[34]:

	model
golf	11341
andere	10559
3er	8360
polo	4865
corsa	4639
...	...
serie_3	2
samara	2
serie_2	2
serie_1	1
discovery_sport	1

250 rows × 1 columns

In [35]:
```
car_models.loc[car_models['model'] > 1000]
```

Out[35]:

	model
golf	11341
andere	10559
3er	8360
polo	4865
corsa	4639
a4	4202
astra	4148
passat	4052
5er	3666
c_klasse	3654
e_klasse	3144
a6	2518
a3	2517
focus	2362
transporter	2253
fiesta	2185
2_reihe	1980
twingo	1738
a_klasse	1609
1er	1576
vectra	1570
fortwo	1505
mondeo	1472
3_reihe	1383
touran	1367
clio	1306
punto	1201
zafira	1153
megane	1131
lupo	1021
ibiza	1018
ka	1002

In [36]:
```
car_models.loc[(car_models['model'] >= 75) & (car_models['model'] <= 1000)]
```

Out[36]:

	model
x_reihe	989
octavia	910
cooper	895
fabia	872
clk	816
...	...
ypsilon	79
jimny	79
cc	78
lancer	77
g_klasse	76

155 rows × 1 columns

In [37]:
```
df = df[df.powerPS >= 50]
```

In [38]:
```
df.describe()
```

Out[38]:

	price	yearOfRegistration	powerPS	kilometer	monthOfRegistration	notRe
count	1.289660e+05	128966.000000	128966.000000	128966.000000	128966.000000	
mean	7.599813e+03	2003.255641	131.896857	125115.922026	6.134725	
std	2.828041e+05	6.247818	177.054593	39289.463954	3.502043	
min	0.000000e+00	1934.000000	50.000000	5000.000000	0.000000	
25%	1.500000e+03	1999.000000	86.000000	100000.000000	3.000000	
50%	3.700000e+03	2004.000000	116.000000	150000.000000	6.000000	
75%	8.450000e+03	2008.000000	155.000000	150000.000000	9.000000	
max	9.900000e+07	2018.000000	20000.000000	150000.000000	12.000000	

In [39]:
```
df.loc[df['powerPS'] >= 1000]
```

Out[39]:

	dateCrawled	name	offerType	price	abtest	vehicleT
514	2016-03-28 12:47:00	opel_vectra	Angebot	1450	test	limou
4576	2016-03-18 15:42:00	Golf_2._CL_Bj_1989	Angebot	1800	control	limou
5785	2016-03-05 16:53:00	A140__Als_Bastler_Fahrzeug_nicht_fahrbereit_	Angebot	850	control	kleinwa
6116	2016-03-20 16:51:00	Verkaufe_meinen_bmw_525d	Angebot	6000	test	kc
7551	2016-03-11 13:58:00	Audi_A4_Quatrro_1.9_TDI	Angebot	900	test	kc
...	...	...	...	...	...	
165025	2016-03-16 21:48:00	Golf_3/3_Tueren	Angebot	850	control	limou
165133	2016-04-04 12:38:00	Opel_Corsa_1.7_cdti	Angebot	3000	control	cc
165806	2016-03-13 11:51:00	Top_gepflegter_Scenic_von_Privat	Angebot	7500	control	limou
165894	2016-04-04 18:53:00	Baster_Fahrzeug	Angebot	250	test	kleinwa
169007	2016-03-07 21:36:00	!!!!!_Opel_zafira_2.0_dti_16_v_comfort_verkauf...	Angebot	1500	control	

94 rows × 20 columns

In [40]:
```
df.loc[df['price'] >= 5000]
```

Out[40]:

	dateCrawled	name	offerType	price	abtes
0	2016-03-26 14:57:00	BMW_320d_DPFXenon_Tempomat_Sitzheizung_PDC_Kli...	Angebot	10499	contr
4	2016-03-26 21:36:00	Ford_Focus_Turnier_1.6_TDCi_DPF_Trend	Angebot	7500	contr
5	2016-03-15 19:45:00	Audi_Q5_3.0_TDI_quattro_S_tronic	Angebot	34800	contr
9	2016-04-01 12:53:00	Renault_Trafic_2.5_dCi_Generation_Expression	Angebot	5200	tes
10	2016-03-20 20:56:00	BMW_Z4_roadster_2.2i	Angebot	8700	tes
...	...	...	...	...	.
171510	2016-03-06 21:11:00	Mercedes_benz_e_klasse_avangarde_220_cdi_grune...	Angebot	6500	tes
171512	2016-03-25 18:48:00	Mercedes_Benz_E_400_CDI_Avantgarde	Angebot	5000	tes
171517	2016-03-28 13:48:00	Volkswagen_Golf_2.0_TDI_DPF_Team	Angebot	7900	tes
171525	2016-03-19 18:57:00	Volkswagen_Multivan_T4_TDI_7DC_UY2	Angebot	9200	tes
171527	2016-03-07 19:39:00	BMW_M135i_vollausgestattet_NP_52.720____Euro	Angebot	28990	contr

52093 rows × 20 columns

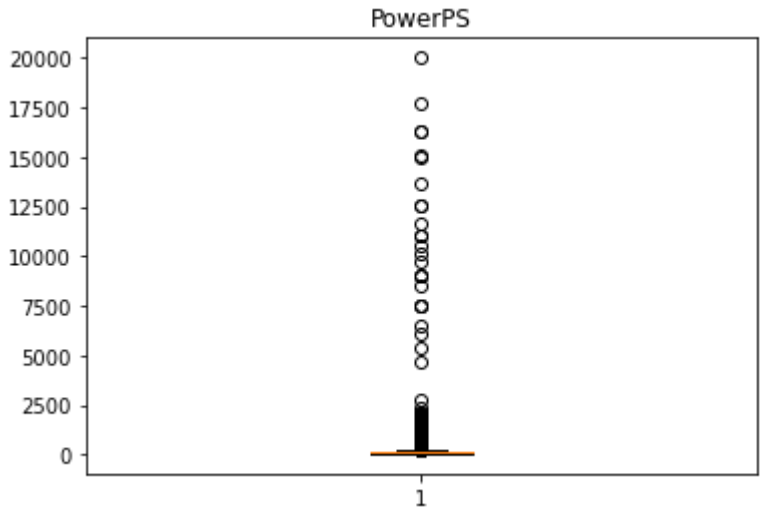
In [41]:
```
df.loc[(df['price'] >= 20000) & (df['status'] == 0)]
```

Out[41]:

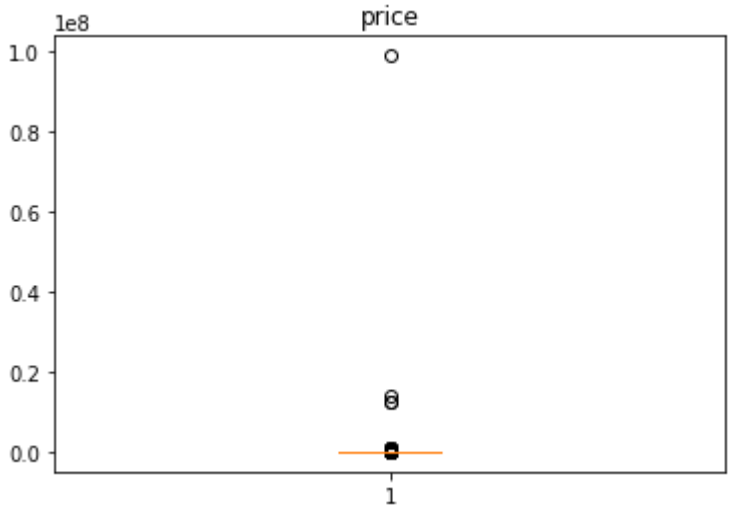
	dateCrawled	name	offerType	price	abte
5	2016-03-15 19:45:00	Audi_Q5_3.0_TDI_quattro_S_tronic	Angebot	34800	cont
24	2016-03-15 21:50:00	Audi_A5_2.0_TDI_Sportback_DPF_multitronic	Angebot	23990	cont
46	2016-03-19 11:56:00	Porsche_Cayenne_Diesel_SportDesign_21"_Luft_1....	Angebot	56800	te
53	2016-03-17 14:47:00	CLA_180_Urban_NEUWAGEN!_20_km_"Neupreis_30800?"	Angebot	24400	cont
65	2016-03-10 12:50:00	Audi_A6_Avant_2.0_TDI_Ultra_DPF	Angebot	31500	te
...	...	...	...	...	
171435	2016-03-31 23:40:00	BMW_525d_Sport_Aut.	Angebot	21499	cont
171477	2016-03-07 16:50:00	BMW_525d_Touring_Sport_Aut.	Angebot	23900	cont
171483	2016-03-07 09:54:00	Mercedes_Benz_E_200_CDI_DPF_BlueEFFICIENCY_7G_...	Angebot	20500	cont
171500	2016-03-21 23:40:00	Volkswagen_Golf_1.4_TSI_BlueMotion_Technology_...	Angebot	20400	te
171527	2016-03-07 19:39:00	BMW_M135i_vollausgestattet_NP_52.720____Euro	Angebot	28990	cont

6567 rows × 20 columns

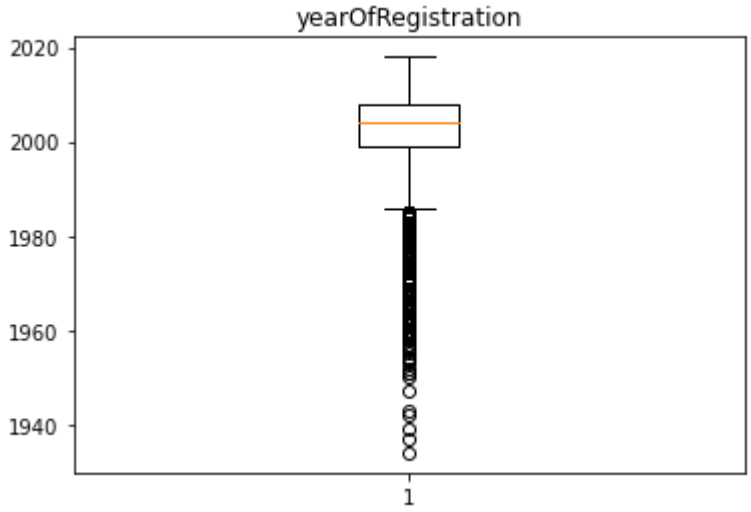
In [42]:
```
plt.title('PowerPS')
plt.boxplot(df.powerPS)
plt.show()
```



In [43]:
```
plt.title('price')
plt.boxplot(df.price)
plt.show()
```

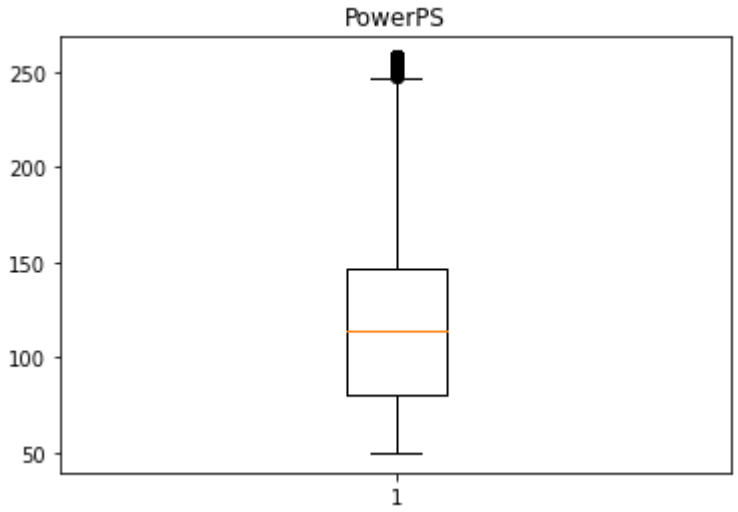


In [44]:
```
plt.title('yearOfRegistration')
plt.boxplot(df.yearOfRegistration)
plt.show()
```

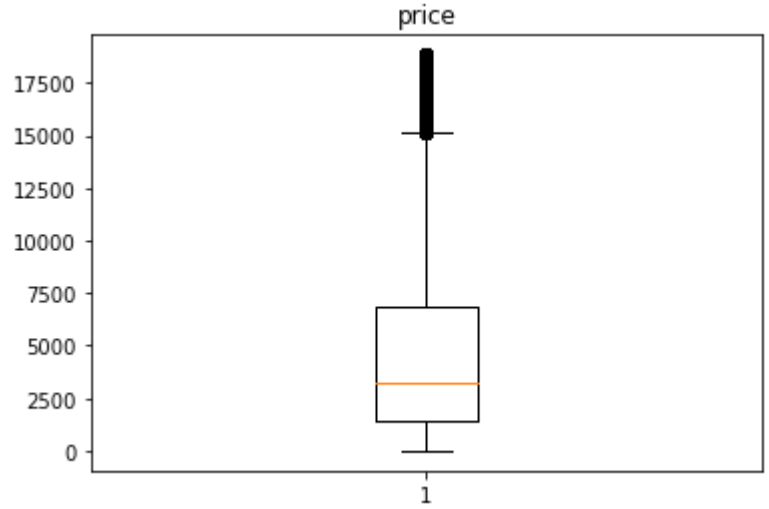


In [45]:
```
cols = ['powerPS', 'price', 'yearOfRegistration']

Q1 = df[cols].quantile(0.25)
Q3 = df[cols].quantile(0.75)
IQR = Q3 - Q1

df = df[~((df[cols] < (Q1 - 1.5 * IQR)) |(df[cols] > (Q3 + 1.5 * IQR))).any(axis=
```

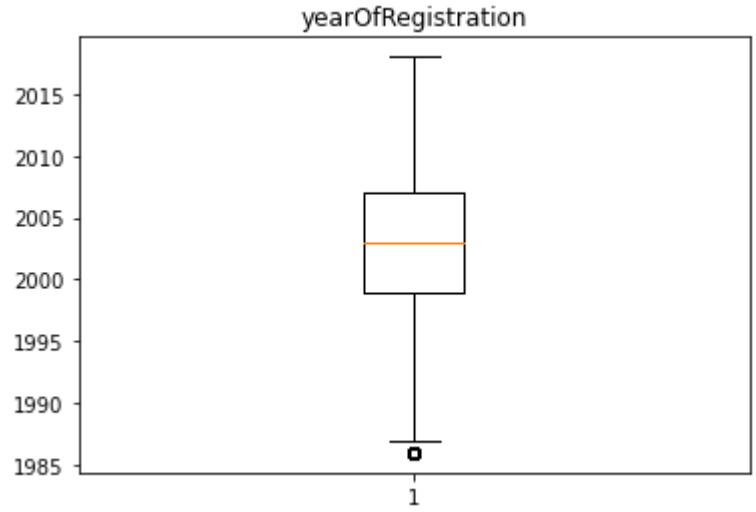
In [46]:
```
plt.title('PowerPS')
plt.boxplot(df.powerPS)
plt.show()
```



In [47]:
```
plt.title('price')
plt.boxplot(df.price)
plt.show()
```



In [48]:
```
plt.title('yearOfRegistration')
plt.boxplot(df.yearOfRegistration)
plt.show()
```



In [49]:
```
df.loc[df['price'] >= 5000]
```

Out[49]:

	dateCrawled	name	offerType	price	abtes
0	2016-03-26 14:57:00	BMW_320d_DPFXenon_Tempomat_Sitzheizung_PDC_Kli...	Angebot	10499	contr
4	2016-03-26 21:36:00	Ford_Focus_Turnier_1.6_TDCi_DPF_Trend	Angebot	7500	contr
9	2016-04-01 12:53:00	Renault_Trafic_2.5_dCi_Generation_Expression	Angebot	5200	tes
10	2016-03-20 20:56:00	BMW_Z4_roadster_2.2i	Angebot	8700	tes
11	2016-03-16 13:50:00	Ford_C_max_1_6_EcoBost_Titanium_Start_Stop_Top...	Angebot	13190	contr
...	...	...	...	...	.
171506	2016-03-20 18:47:00	Volkswagen_Golf_1.9_TDI_DPF_Goal_Rentnerfzg._1...	Angebot	5900	tes
171507	2016-03-16 17:06:00	Gepflegter_Audi_a4_2.0tdi_xenon_ahk_klima_sche...	Angebot	5999	tes
171510	2016-03-06 21:11:00	Mercedes_benz_e_klasse_avangarde_220_cdi_grune...	Angebot	6500	tes
171517	2016-03-28 13:48:00	Volkswagen_Golf_2.0_TDI_DPF_Team	Angebot	7900	tes
171525	2016-03-19 18:57:00	Volkswagen_Multivan_T4_TDI_7DC_UY2	Angebot	9200	tes

41125 rows × 20 columns

In [50]:
```
df
```

Out[50]:

	dateCrawled	name	offerType	price	abtes
0	2016-03-26 14:57:00	BMW_320d_DPFXenon_Tempomat_Sitzheizung_PDC_Kli...	Angebot	10499	contr
2	2016-03-19 18:56:00	VW_LUPO._1.1l_T?_3/18	Angebot	2750	tes
3	2016-03-30 12:51:00	Opel_Astra_Caravan_Sport_sehr_gepflegt	Angebot	3500	contr
4	2016-03-26 21:36:00	Ford_Focus_Turnier_1.6_TDCi_DPF_Trend	Angebot	7500	contr
6	2016-03-07 22:48:00	Alfa_Romeo_147	Angebot	1050	tes
...	...	...	...	...	.
171517	2016-03-28 13:48:00	Volkswagen_Golf_2.0_TDI_DPF_Team	Angebot	7900	tes
171520	2016-03-19 19:53:00	turbo_defekt	Angebot	3200	contr
171524	2016-03-05 19:56:00	Smart_smart_leistungssteigerung_100ps	Angebot	1199	tes
171525	2016-03-19 18:57:00	Volkswagen_Multivan_T4_TDI_7DC_UY2	Angebot	9200	tes
171526	2016-03-20 19:41:00	VW_Golf_Kombi_1_9l_TDI	Angebot	3400	tes

116931 rows × 20 columns

In [51]:
```
df.name.value_counts()
```

Out[51]:
```
BMW_318i                                                        296
Volkswagen_Golf_1.4                                             261
BMW_320i                                                        229
BMW_316i                                                        228
Ford_Fiesta                                                     198
                                                               ...
Peugeot_307_CC_140_JBL_Cabrio                                     1
BMW_X5_3.0_d_Gelegenheit!_Vollausstattung___wenig_Kilometer!      1
Peugeot_206cc_T?_07_2016                                          1
VW_Golf_2_GT_G60_Renner__Slalom__Viertel_Meile                    1
VW_Golf_Kombi_1_9l_TDI                                            1
Name: name, Length: 75522, dtype: int64
```

In [52]:
```
df.describe()
```

Out[52]:

	price	yearOfRegistration	powerPS	kilometer	monthOfRegistration	notR
count	116931.000000	116931.000000	116931.000000	116931.000000	116931.000000	
mean	4728.572620	2003.035577	118.587817	128653.949765	6.132078	
std	4333.565384	5.359477	44.888983	36147.280310	3.512960	
min	0.000000	1986.000000	50.000000	5000.000000	0.000000	
25%	1399.000000	1999.000000	80.000000	125000.000000	3.000000	
50%	3200.000000	2003.000000	114.000000	150000.000000	6.000000	
75%	6900.000000	2007.000000	147.000000	150000.000000	9.000000	
max	18850.000000	2018.000000	258.000000	150000.000000	12.000000	

In [53]:
```
df.loc[df['yearOfRegistration'] <= 1990]
```

Out[53]:

	dateCrawled	name	offerType	price	abtest	v
307	2016-04-04 02:36:00	Audi_80_T?_2017	Angebot	500	control	
377	2016-03-21 02:36:00	BMW_535i_e34_mit_vielen_Extras	Angebot	3800	test	
489	2016-04-01 20:57:00	Audi_100_quattro__2.3__Tuev_09_2017__Anhaenger...	Angebot	1200	test	
520	2016-03-16 14:54:00	Mercedes_Benz_CE_300	Angebot	3890	test	
603	2016-03-12 19:00:00	Volkswagen_Golf_2_Rat_style_T?_NEU_top_zustand	Angebot	1450	test	
...	...	...	...	...	...	
171376	2016-03-05 21:57:00	Volkswagen_Golf_2_VR6_alles_Eingetragen	Angebot	4250	test	
171391	2016-03-17 13:38:00	BMW_730i	Angebot	2500	control	
171439	2016-04-05 07:37:00	Volkswagen_T3_DoKa_Pritsche_Transporter_1_6_TD	Angebot	1950	test	
171459	2016-03-05 16:44:00	Audi_80_quattro	Angebot	850	control	
171468	2016-04-04 19:45:00	VW_Golf_II	Angebot	1550	test	

1672 rows × 20 columns

In [54]:
```
for col in df:
    print(col + " ")
    print(df[col].unique())
```

```
dateCrawled
['2016-03-26T14:57:00.000000000' '2016-03-19T18:56:00.000000000'
 '2016-03-30T12:51:00.000000000' ... '2016-03-26T08:52:00.000000000'
 '2016-04-05T07:37:00.000000000' '2016-03-06T21:11:00.000000000']
name
['BMW_320d_DPFXenon_Tempomat_Sitzheizung_PDC_Klima_MFL'
 'VW_LUPO._1.1l_T?_3/18' 'Opel_Astra_Caravan_Sport_sehr_gepflegt' ...
 'turbo_defekt' 'Smart_smart_leistungssteigerung_100ps'
 'VW_Golf_Kombi_1_9l_TDI']
offerType
['Angebot' 'Gesuch']
price
[10499  2750  3500 ... 10985 14989 15190]
abtest
['control' 'test']
vehicleType
['limousine' 'kleinwagen' 'kombi' 'bus' 'cabrio' 'coupe' 'andere' 'suv']
yearOfRegistration
[2006 1999 2004 2012 2002 1998 2005 2011 2001 2007 2003 2008 2013 2009
 1994 2010 2000 1993 1997 2014 1995 1991 1992 1996 1989 1990 1987 1988
 1986 2016 2015 2017 2018]
gearbox
['manual' 'automatik']
powerPS
[163  50 125 116 105 115 135 170 182  86 140 232  64 136 126  95 165 109
 150 122  75  90 101 131 132  60 143  69 102 156  54  85  73  68 118  74
 120 110 103 231  72  92  97 204 160  65  80 218 113 145 215 224 239 177
 174 241 173  58 235 167 193 129 250 112 192 225 107  77 185  98 152  88
 210 220 114  59 128 104  82 100 211 190 141  55 179  84  71 133 197  76
 200 155 130 228 245  67 111 147 184 205  91  94 175  61 180  63  78 117
 139  79 230 209 256 144 124 188  87  83 226 137 121 240  66 154 166  96
  70 169  56 191 176 108 194 171  99 178 189 119 146  52  53 106  57  51
 207 195  81 151 158 233 172 196 219  62 258 252 238 217 202 148 186 127
 142 213 201 161 203  93 236 149 223 162 157  89 198 159 254 164 247 212
 168 123 243 181 253 214 249 134 199 138 242 183 234 246 222 237 187 244
 255 227 216 229 153 248 221 208 206 257]
model
['3er' 'lupo' 'astra' 'focus' '147' 'espace' 'andere' 'z_reihe' 'c_max'
 'yaris' 'a3' 'a6' 'polo' '2_reihe' 'auris' 'passat' 'golf' 'ceed' '1er'
 'accord' '3_reihe' 'corsa' 'c_klasse' 'clio' 'meriva' 'zafira' 'caddy'
 'a4' 'fox' 'one' 'transporter' 'laguna' 'touran' 'fiesta' 'tiguan'
 'berlingo' 'stilo' '5er' 'mondeo' 'corolla' 'yeti' 'cr_reihe' 'qashqai'
 'micra' 'aygo' 'eos' 'a1' 'fortwo' 'jimny' 'galaxy' 'e_klasse' 'slk'
 'megane' 'omega' 'c1' 'civic' '90' 'mx_reihe' 'm_klasse' 'a8' 'cooper'
 'q5' 'x_type' 'fabia' 'transit' 'rio' 'punto' 'a_klasse' 'panda' 'scenic'
 'tucson' 'tt' 'alhambra' 'kangoo' '80' 'twingo' 'vectra' 'c4' 'seicento'
 'c3' 'scirocco' 'ibiza' 'primera' 'rav' 'forfour' 'ka' 'picanto' 'niva'
 'beetle' 'phaeton' 's_klasse' 'clk' 'a5' 'tigra' 'sharan' '100' 'voyager'
 '7er' 'carisma' 'logan' 'sprinter' '6_reihe' '1_reihe' 'calibra'
 'octavia' 'modus' 'a2' 's_max' 'boxster' 'outlander' 'i_reihe' 'leon'
 '500' 'fusion' 'lancer' 'freelander' 'verso' 'grand' 'swift' 'x_reihe'
 'ypsilon' 'arosa' 'vivaro' 'sl' 'vito' 'c5' 's60' 'lybra' 'xc_reihe'
```

```
 'almera' 'sorento' 'signum' '900' 'v70' 'range_rover' 'b_klasse' 'superb'
 'colt' 'getz' 'escort' 'c_reihe' 'materia' 'v50' 'viano' '5_reihe'
 'cuore' 'roomster' 'carnival' 'cordoba' 'sportage' 'doblo' '156'
 'x_trail' '850' 'c2' 'v40' 'kuga' 'altea' 'cx_reihe' '4_reihe' 'jazz'
 'pajero' 'spark' 'bora' 'sandero' 'bravo' 'juke' 'avensis' 'rx_reihe'
 'insignia' 'duster' 'toledo' 'cayenne' 'galant' 'agila' 'combo' 'santa'
 'matiz' 'terios' 'defender' 'sirion' 'mustang' 'antara' 'v_klasse'
 'jetta' 'touareg' '159' 'justy' 'kalos' 'impreza' 'clubman' 'up' 'croma'
 'note' 'lodgy' 'captiva' 'kadett' 'cc' 'g_klasse' 'aveo' 'citigo'
 's_type' 'ptcruiser' 'legacy' 'q7' 'mii' '300c' '145' 'cl' 'r19' 'ducato'
 'spider' '6er' 'forester' 'nubira' 'discovery' 'roadster'
 'range_rover_sport' 'lanos' 'exeo' '9000' 'samara' 'navara' 'glk'
 'crossfire' 'cherokee' 'delta' 'wrangler' 'i3' 'kappa' 'move' 'musa'
 'm_reihe' 'kalina' 'charade' 'amarok' '200' 'gl' 'v60' 'b_max' 'q3'
 'elefantino' '911' 'kaefer']
kilometer
[ 90000  40000 150000  50000 125000  70000  80000  30000  60000 100000
  20000   5000  10000]
monthOfRegistration
[ 1 10  8  6  7  3 12  5  4  2  9  0 11]
fuelType
['diesel' 'benzin' 'lpg' 'cng' 'hybrid' 'andere' 'elektro']
brand
['bmw' 'volkswagen' 'opel' 'ford' 'alfa_romeo' 'renault' 'toyota' 'audi'
 'peugeot' 'kia' 'honda' 'saab' 'mercedes_benz' 'fiat' 'mini' 'citroen'
 'volvo' 'skoda' 'nissan' 'mazda' 'smart' 'suzuki' 'jaguar' 'hyundai'
 'seat' 'rover' 'lada' 'chrysler' 'jeep' 'mitsubishi' 'dacia' 'porsche'
 'land_rover' 'lancia' 'chevrolet' 'daewoo' 'daihatsu' 'subaru' 'trabant']
notRepairedDamage
[0 1]
dateCreated
['2016-03-26T00:00:00.000000000' '2016-03-19T00:00:00.000000000'
 '2016-03-30T00:00:00.000000000' '2016-03-07T00:00:00.000000000'
 '2016-04-01T00:00:00.000000000' '2016-03-20T00:00:00.000000000'
 '2016-03-16T00:00:00.000000000' '2016-03-08T00:00:00.000000000'
 '2016-03-11T00:00:00.000000000' '2016-03-29T00:00:00.000000000'
 '2016-03-31T00:00:00.000000000' '2016-03-05T00:00:00.000000000'
 '2016-03-12T00:00:00.000000000' '2016-04-02T00:00:00.000000000'
 '2016-04-04T00:00:00.000000000' '2016-03-22T00:00:00.000000000'
 '2016-03-17T00:00:00.000000000' '2016-03-10T00:00:00.000000000'
 '2016-03-25T00:00:00.000000000' '2016-03-15T00:00:00.000000000'
 '2016-03-21T00:00:00.000000000' '2016-04-03T00:00:00.000000000'
 '2016-03-28T00:00:00.000000000' '2016-03-27T00:00:00.000000000'
 '2016-03-14T00:00:00.000000000' '2016-03-09T00:00:00.000000000'
 '2016-03-06T00:00:00.000000000' '2016-03-18T00:00:00.000000000'
 '2016-04-05T00:00:00.000000000' '2016-03-13T00:00:00.000000000'
 '2016-03-23T00:00:00.000000000' '2016-03-24T00:00:00.000000000'
 '2016-04-07T00:00:00.000000000' '2016-04-06T00:00:00.000000000'
 '2016-02-14T00:00:00.000000000' '2016-03-02T00:00:00.000000000'
 '2016-02-24T00:00:00.000000000' '2016-01-13T00:00:00.000000000'
 '2016-03-04T00:00:00.000000000' '2016-02-27T00:00:00.000000000'
 '2016-02-05T00:00:00.000000000' '2016-03-01T00:00:00.000000000'
 '2016-03-03T00:00:00.000000000' '2016-02-18T00:00:00.000000000'
 '2016-01-30T00:00:00.000000000' '2016-02-29T00:00:00.000000000'
 '2016-02-13T00:00:00.000000000' '2016-02-25T00:00:00.000000000'
 '2016-01-28T00:00:00.000000000' '2016-01-19T00:00:00.000000000'
 '2016-02-20T00:00:00.000000000' '2016-02-28T00:00:00.000000000'
```

```
 '2016-02-17T00:00:00.000000000' '2016-02-21T00:00:00.000000000'
 '2016-02-12T00:00:00.000000000' '2016-02-22T00:00:00.000000000'
 '2016-02-26T00:00:00.000000000' '2015-09-04T00:00:00.000000000'
 '2015-12-05T00:00:00.000000000' '2016-01-31T00:00:00.000000000'
 '2016-02-19T00:00:00.000000000' '2016-02-02T00:00:00.000000000'
 '2016-02-10T00:00:00.000000000' '2016-01-25T00:00:00.000000000'
 '2016-02-07T00:00:00.000000000' '2015-12-30T00:00:00.000000000'
 '2016-01-23T00:00:00.000000000' '2016-02-16T00:00:00.000000000'
 '2016-02-09T00:00:00.000000000' '2016-01-24T00:00:00.000000000'
 '2015-12-17T00:00:00.000000000' '2016-02-01T00:00:00.000000000'
 '2015-11-10T00:00:00.000000000' '2016-01-26T00:00:00.000000000'
 '2016-01-07T00:00:00.000000000' '2016-01-10T00:00:00.000000000'
 '2015-11-23T00:00:00.000000000' '2016-01-06T00:00:00.000000000'
 '2016-01-29T00:00:00.000000000' '2016-01-27T00:00:00.000000000'
 '2015-09-09T00:00:00.000000000' '2016-02-04T00:00:00.000000000'
 '2016-01-02T00:00:00.000000000']
postalCode
['76661' '53757' '4329' ... '54585' '96269' '84181']
lastSeen
['2016-04-06T02:46:00.000000000' '2016-03-23T08:15:00.000000000'
 '2016-04-01T12:50:00.000000000' ... '2016-04-02T12:33:00.000000000'
 '2016-03-08T20:43:00.000000000' '2016-03-18T11:30:00.000000000']
days
[10  3  1  6  4  5 17 13  0 28 21  2  8 30 25  7 19 24 20 16 29 14 18 12
  9 26 22 11 23 15 32 27 31 33]
status
[0 1]
```

In [55]:
```
df.describe()
```

Out[55]:

	price	yearOfRegistration	powerPS	kilometer	monthOfRegistration	notR
count	116931.000000	116931.000000	116931.000000	116931.000000	116931.000000	
mean	4728.572620	2003.035577	118.587817	128653.949765	6.132078	
std	4333.565384	5.359477	44.888983	36147.280310	3.512960	
min	0.000000	1986.000000	50.000000	5000.000000	0.000000	
25%	1399.000000	1999.000000	80.000000	125000.000000	3.000000	
50%	3200.000000	2003.000000	114.000000	150000.000000	6.000000	
75%	6900.000000	2007.000000	147.000000	150000.000000	9.000000	
max	18850.000000	2018.000000	258.000000	150000.000000	12.000000	

In [56]:
```
df
```

Out[56]:

	dateCrawled	name	offerType	price	abtes
0	2016-03-26 14:57:00	BMW_320d_DPFXenon_Tempomat_Sitzheizung_PDC_Kli...	Angebot	10499	contr
2	2016-03-19 18:56:00	VW_LUPO._1.1l_T?_3/18	Angebot	2750	tes
3	2016-03-30 12:51:00	Opel_Astra_Caravan_Sport_sehr_gepflegt	Angebot	3500	contr
4	2016-03-26 21:36:00	Ford_Focus_Turnier_1.6_TDCi_DPF_Trend	Angebot	7500	contr
6	2016-03-07 22:48:00	Alfa_Romeo_147	Angebot	1050	tes
...	...	...	...	...	.
171517	2016-03-28 13:48:00	Volkswagen_Golf_2.0_TDI_DPF_Team	Angebot	7900	tes
171520	2016-03-19 19:53:00	turbo_defekt	Angebot	3200	contr
171524	2016-03-05 19:56:00	Smart_smart_leistungssteigerung_100ps	Angebot	1199	tes
171525	2016-03-19 18:57:00	Volkswagen_Multivan_T4_TDI_7DC_UY2	Angebot	9200	tes
171526	2016-03-20 19:41:00	VW_Golf_Kombi_1_9l_TDI	Angebot	3400	tes

116931 rows × 20 columns

In [76]:
```
df_copy = df.copy()
```

In [77]:
```
df_copy = df_copy.drop(colgenres = np.unique(vg_df['Genre'])
genresumns=['dateCrawled', 'dateCreated', 'lastSeen', 'lastSeen', 'name', 'days',
```

In [78]:
```
df_copy
```

Out[78]:

	offerType	price	abtest	vehicleType	yearOfRegistration	gearbox	powerPS	model
0	Angebot	10499	control	limousine	2006	manual	163	3er
2	Angebot	2750	test	kleinwagen	1999	manual	50	lupo
3	Angebot	3500	control	kombi	2004	manual	125	astra
4	Angebot	7500	control	kombi	2012	manual	116	focus
6	Angebot	1050	test	kleinwagen	2002	manual	105	147
...	...	...	...	...	...	...	...	...
171517	Angebot	7900	test	limousine	2010	manual	140	golf
171520	Angebot	3200	control	limousine	2004	manual	225	leon
171524	Angebot	1199	test	cabrio	2000	automatik	101	fortwo
171525	Angebot	9200	test	bus	1996	manual	102	transporter
171526	Angebot	3400	test	kombi	2002	manual	100	golf

116931 rows × 14 columns

In [69]:
```
car_models_df_copy = df_copy.model.value_counts()
```

In [70]:
```
car_models_df_copy = pd.DataFrame(car_models_df_copy)
```

In [71]:
```
car_models_df_copy
```

Out[71]:

	model
golf	10301
andere	7605
3er	7425
polo	4141
astra	3858
...	...
i3	2
samara	2
elefantino	2
911	1
kaefer	1

244 rows × 1 columns

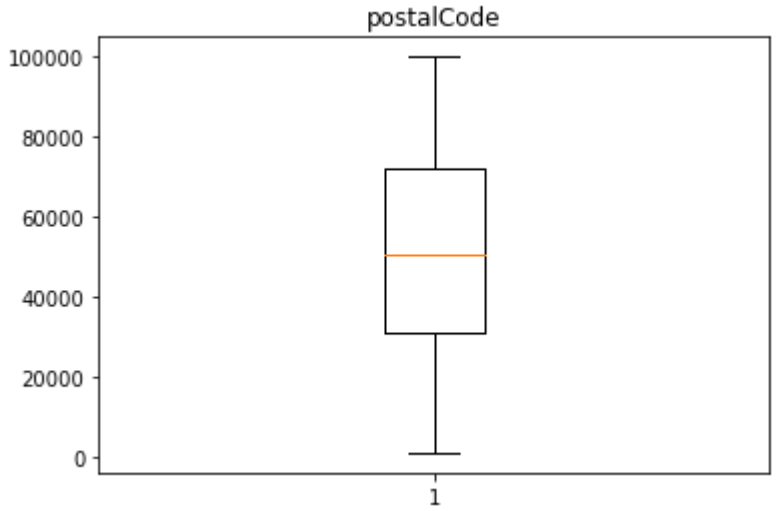
In [72]:
```
df["postalCode"] = pd.to_numeric(df["postalCode"])
```

In [74]:
```
df.describe()
```

Out[74]:

	price	yearOfRegistration	powerPS	kilometer	monthOfRegistration	notR
count	116931.000000	116931.000000	116931.000000	116931.000000	116931.000000	
mean	4728.572620	2003.035577	118.587817	128653.949765	6.132078	
std	4333.565384	5.359477	44.888983	36147.280310	3.512960	
min	0.000000	1986.000000	50.000000	5000.000000	0.000000	
25%	1399.000000	1999.000000	80.000000	125000.000000	3.000000	
50%	3200.000000	2003.000000	114.000000	150000.000000	6.000000	
75%	6900.000000	2007.000000	147.000000	150000.000000	9.000000	
max	18850.000000	2018.000000	258.000000	150000.000000	12.000000	

In [75]:
```
plt.title('postalCode')
plt.boxplot(df.postalCode)
plt.show()
```



In [81]:
```
#df_copy.to_csv('Automotive_2_clean_data.csv', index = False)
```

In [82]:
```
df_copy
```

Out[82]:

	offerType	price	abtest	vehicleType	yearOfRegistration	gearbox	powerPS	model
0	Angebot	10499	control	limousine	2006	manual	163	3er
2	Angebot	2750	test	kleinwagen	1999	manual	50	lupo
3	Angebot	3500	control	kombi	2004	manual	125	astra
4	Angebot	7500	control	kombi	2012	manual	116	focus
6	Angebot	1050	test	kleinwagen	2002	manual	105	147
...	...	...	...	...	...	...	...	...
171517	Angebot	7900	test	limousine	2010	manual	140	golf
171520	Angebot	3200	control	limousine	2004	manual	225	leon
171524	Angebot	1199	test	cabrio	2000	automatik	101	fortwo
171525	Angebot	9200	test	bus	1996	manual	102	transporter
171526	Angebot	3400	test	kombi	2002	manual	100	golf

116931 rows × 14 columns

In [ ]:
```

```