Import Libraries

```
import pandas as pd

import numpy as np

import matplotlib as plt

from sklearn.preprocessing import LabelEncoder

import pickle
```

Import Dataset

```
df=pd.read_csv("E:/IBM/Collect Dataset/Dataset.csv", header=0, sep=',', encoding='Latin1',)
```

Read Dataset

```
df.head()
```

```
dateCrawled     name    seller  offerType       price   abtest  vehicleType     yearOfRegistration
        gearbox         powerPS         model   kilometer       monthOfRegistration     fuelType
        brand   notRepairedDamage       dateCreated     nrOfPictures    postalCode      lastSeen

0       2016-03-24 11:52:17     Golf_3_1.6      privat  Angebot         480     test    NaN     1993
        manuell         0       golf    150000 0        benzin  volkswagen      NaN     2016-03-24
00:00:00        0       70435   2016-04-07 03:16:57

1       2016-03-24 10:58:45     A5_Sportback_2.7_Tdi    privat  Angebot         18300   test    coupe
        2011    manuell         190     NaN     125000 5        diesel  audi    ja      2016-03-24
00:00:00        0       66954   2016-04-07 01:46:50

2       2016-03-14 12:52:21     Jeep_Grand_Cherokee_"Overland"  privat  Angebot         9800
        test    suv     2004    automatik       163     grand   125000 8        diesel  jeep    NaN
        2016-03-14 00:00:00     0       90480   2016-04-05 12:47:46

3       2016-03-17 16:54:04     GOLF_4_1_4__3TÜRER privat       Angebot         1500    test
        kleinwagen      2001    manuell         75      golf    150000 6        benzin  volkswagen
        nein    2016-03-17 00:00:00     0       91074   2016-03-17 17:40:17

4       2016-03-31 17:25:20     Skoda_Fabia_1.4_TDI_PD_Classic  privat  Angebot         3600
        test    kleinwagen      2008    manuell         69      fabia   90000   7       diesel  skoda
        nein    2016-03-31 00:00:00     0       60437   2016-04-06 10:17:21
```

```
df.shape
```

```
(371528, 20)
```

```
print(df.seller.value_counts())
```

```
privat       371525

gewerblich        3
```

```
Name: seller, dtype: int64
```

```
df[df.seller != 'gewerblich']
```

```
dateCrawled     name    seller  offerType       price   abtest  vehicleType     yearOfRegistration
        gearbox         powerPS         model   kilometer       monthOfRegistration     fuelType
        brand   notRepairedDamage       dateCreated     nrOfPictures    postalCode      lastSeen

0       2016-03-24 11:52:17     Golf_3_1.6      privat  Angebot         480     test    NaN     1993
        manuell         0       golf    150000 0        benzin  volkswagen      NaN     2016-03-24
00:00:00        0       70435   2016-04-07 03:16:57

1       2016-03-24 10:58:45     A5_Sportback_2.7_Tdi    privat  Angebot         18300   test    coupe
        2011    manuell         190     NaN     125000 5        diesel  audi    ja      2016-03-24
00:00:00        0       66954   2016-04-07 01:46:50

2       2016-03-14 12:52:21     Jeep_Grand_Cherokee_"Overland"  privat  Angebot         9800
        test    suv     2004    automatik       163     grand   125000 8        diesel  jeep    NaN
        2016-03-14 00:00:00     0       90480   2016-04-05 12:47:46

3       2016-03-17 16:54:04     GOLF_4_1_4__3TÜRER privat  Angebot         1500    test
        kleinwagen      2001    manuell         75      golf    150000 6        benzin  volkswagen
        nein    2016-03-17 00:00:00     0       91074   2016-03-17 17:40:17

4       2016-03-31 17:25:20     Skoda_Fabia_1.4_TDI_PD_Classic  privat  Angebot         3600
        test    kleinwagen      2008    manuell         69      fabia   90000   7       diesel  skoda
        nein    2016-03-31 00:00:00     0       60437   2016-04-06 10:17:21

...     ...     ...     ...     ...     ...     ...     ...     ...     ...     ...     ...     ...
        ...     ...     ...     ...     ...     ...     ...     ...

371523 2016-03-14 17:48:27      Suche_t4___vito_ab_6_sitze      privat  Angebot         2200    test
        NaN     2005    NaN     0       NaN     20000   1       NaN     sonstige_autos NaN      2016-
03-14 00:00:00 0        39576   2016-04-06 00:46:52

371524 2016-03-05 19:56:21      Smart_smart_leistungssteigerung_100ps   privat  Angebot
        1199    test    cabrio  2000    automatik       101     fortwo  125000 3        benzin  smart
        nein    2016-03-05 00:00:00     0       26135   2016-03-11 18:17:12

371525 2016-03-19 18:57:12      Volkswagen_Multivan_T4_TDI_7DC_UY2      privat  Angebot
        9200    test    bus     1996    manuell         102     transporter     150000 3        diesel
        volkswagen      nein    2016-03-19 00:00:00     0       87439   2016-04-07 07:15:26

371526 2016-03-20 19:41:08      VW_Golf_Kombi_1_9l_TDI  privat  Angebot         3400    test
        kombi   2002    manuell         100     golf    150000 6        diesel  volkswagen      NaN
        2016-03-20 00:00:00     0       40764   2016-03-24 12:45:21

371527 2016-03-07 19:39:19      BMW_M135i_vollausgestattet_NP_52.720____Euro    privat
        Angebot         28990   control limousine       2013    manuell         320     m_reihe
        50000   8       benzin  bmw     nein    2016-03-07 00:00:00     0       73326   2016-03-22
03:17:10
```

371525 rows × 20 columns

```
df=df.drop('seller',axis=1)
```

Cleaning The Dataset

```
print(df.offerType.value_counts())
```

```
Angebot    371516
Gesuch         12
Name: offerType, dtype: int64
```

```
df[df.offerType != 'Gesuch']
```

```
dateCrawled     name    offerType       price   abtest  vehicleType     yearOfRegistration
        gearbox         powerPS         model   kilometer       monthOfRegistration     fuelType
        brand   notRepairedDamage       dateCreated     nrOfPictures    postalCode      lastSeen

0       2016-03-24 11:52:17     Golf_3_1.6      Angebot         480     test    NaN     1993
        manuell         0       golf    150000 0        benzin  volkswagen      NaN     2016-03-24
00:00:00        0       70435   2016-04-07 03:16:57

1       2016-03-24 10:58:45     A5_Sportback_2.7_Tdi    Angebot         18300   test    coupe   2011
        manuell         190     NaN     125000 5        diesel  audi    ja      2016-03-24 00:00:00
        0       66954   2016-04-07 01:46:50

2       2016-03-14 12:52:21     Jeep_Grand_Cherokee_"Overland"  Angebot         9800    test
        suv     2004    automatik       163     grand   125000 8        diesel  jeep    NaN     2016-
03-14 00:00:00 0        90480   2016-04-05 12:47:46

3       2016-03-17 16:54:04     GOLF_4_1_4__3TÜRER Angebot      1500    test    kleinwagen
        2001    manuell         75      golf    150000 6        benzin  volkswagen      nein    2016-
03-17 00:00:00 0        91074   2016-03-17 17:40:17

4       2016-03-31 17:25:20     Skoda_Fabia_1.4_TDI_PD_Classic  Angebot         3600    test
        kleinwagen      2008    manuell         69      fabia   90000   7       diesel  skoda   nein
        2016-03-31 00:00:00     0       60437   2016-04-06 10:17:21

...     ...     ...     ...     ...     ...     ...     ...     ...     ...     ...     ...     ...
        ...     ...     ...     ...     ...     ...     ...

371523 2016-03-14 17:48:27      Suche_t4___vito_ab_6_sitze      Angebot         2200    test    NaN
        2005    NaN     0       NaN     20000   1       NaN     sonstige_autos  NaN     2016-03-14
00:00:00        0       39576   2016-04-06 00:46:52
```

```
371524 2016-03-05 19:56:21      Smart_smart_leistungssteigerung_100ps          Angebot         1199
        test       cabrio   2000      automatik       101      fortwo  125000 3          benzin  smart    nein
        2016-03-05 00:00:00      0           26135    2016-03-11 18:17:12

371525 2016-03-19 18:57:12      Volkswagen_Multivan_T4_TDI_7DC_UY2           Angebot         9200
        test       bus        1996      manuell          102      transporter       150000 3           diesel
        volkswagen       nein      2016-03-19 00:00:00      0          87439    2016-04-07 07:15:26

371526 2016-03-20 19:41:08      VW_Golf_Kombi_1_9l_TDI       Angebot          3400      test         kombi
        2002      manuell           100       golf       150000 6          diesel    volkswagen       NaN      2016-
03-20 00:00:00  0          40764    2016-03-24 12:45:21

371527 2016-03-07 19:39:19      BMW_M135i_vollausgestattet_NP_52.720____Euro       Angebot
        28990   control  limousine          2013      manuell           320       m_reihe          50000   8
        benzin  bmw      nein      2016-03-07 00:00:00      0          73326    2016-03-22 03:17:10

371516 rows × 19 columns
```

```
df=df.drop('offerType',axis=1)

print(df.shape)

(371528, 1😎

df=df[(df.powerPS > 50) & (df.powerPS < 900)]

print(df.shape)

(319709, 1😎

df = df[(df.yearOfRegistration >= 1950) & (df.yearOfRegistration < 2017)]

print(df.shape)

(309171, 1😎

df.drop(['name', 'abtest', 'dateCrawled', 'nrOfPictures', 'lastSeen',

       'postalCode','dateCreated'], axis='columns',inplace=True)

new_df = df.copy()

new_df = new_df.drop_duplicates ([ 'price', 'vehicleType', 'yearOfRegistration'

                             ,'gearbox', 'powerPS', 'model', 'kilometer', 'monthOfRegistration', 'fuelType'

                             ,'notRepairedDamage'])

new_df.gearbox.replace(('manuell', 'automatik'), ('manual', 'automatic'), inplace=True)

new_df.fuelType.replace(('benzin', 'andere', 'elektro'), ('petrol', 'others', 'electric'), inplace=True)
new_df.vehicleType.replace(('kleinwagen', 'cabrio', 'kombi', 'andere'),
                        ('small car', 'convertible', 'combination', 'others'), inplace=True)
new_df.notRepairedDamage.replace(('ja', 'nein'), ('Yes', 'No'),inplace=True)
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[ 'vehicleType'].fillna (value='not-declared', inplace=True)
new_df['model'].fillna(value='not-declared',inplace=True)
new_df.to_csv("autos_preprocessed.csv")
labels = ['gearbox', 'notRepairedDamage', 'model', 'brand', 'fuelType', 'vehicleType']
mapper = {}
for i in labels:
    mapper[i]=LabelEncoder()
    mapper[i].fit(new_df[i])
    tr = mapper[i].transform(new_df[i])
    np.save(str('classes'+i+ '.npy'), mapper[i].classes_)
    print(i, ":",mapper[i])
    new_df.loc[:, i + '_labels'] = pd.Series (tr, index=new_df.index)
```

```
gearbox : LabelEncoder()
notRepairedDamage : LabelEncoder()
model : LabelEncoder()
brand : LabelEncoder()
fuelType : LabelEncoder()
vehicleType : LabelEncoder()
```

```
labeled=new_df[ ['price'
               ,'yearOfRegistration'
               ,'powerPS'
               ,'kilometer'
```

```
                 ,'monthOfRegistration'
                ]
               + [x+"_labels" for x in labels]]
print(labeled.columns)
Index(['price', 'yearOfRegistration', 'powerPS', 'kilometer',
       'monthOfRegistration', 'gearbox_labels', 'notRepairedDamage_labels',
       'model_labels', 'brand_labels', 'fuelType_labels',
       'vehicleType_labels'],
      dtype='object')
```

Splitting Data Into Independent And Dependent Variables

```
Y = labeled.iloc[:,0].values
X = labeled.iloc[:,1:].values
Y=Y.reshape(-1,1)
from sklearn.model_selection import cross_val_score, train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3, random_state=3)
```