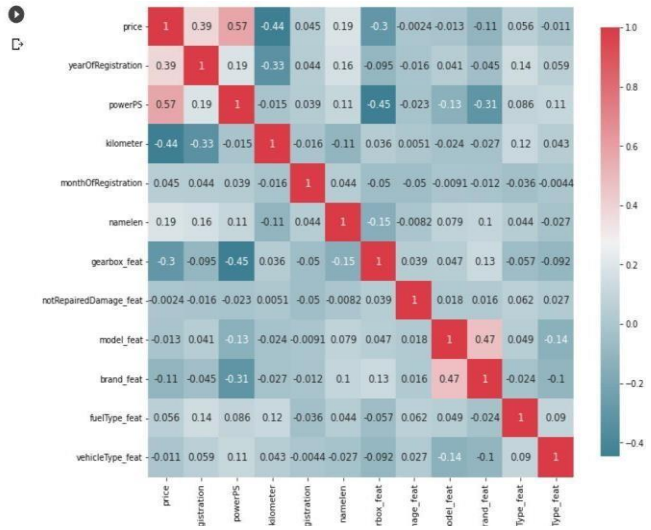
PROJECT DEVELOPMENT PHASE
(DELIVERY OF SPRINT-1)

Date	**19 November 2022**
Team ID	**PNT2022TMID39573**
Project Name	**Car Resale Value prediction**
Maximum marks	**4 Marks**

Import Library and load the data set:

```
df = pd.read_csv('/content/drive/MyDrive/Imarticus/autos.csv', sep=',', header=0, encoding='cp1252')
#df = pd.read_csv('autos.csv.gz', sep=',', header=0, compression='gzip',encoding='cp1252')
df.sample(10)
```

	dateCrawled	name	seller	offerType	price	abtest	vehicleType	yearOfRegistration	gearbox	powerPS	model	kilometer	monthOfRegi
35533	2016-04-01 16:52:24	Peugeot_206_5Tuerer_Klima_El_Fenster_2_Hand_8f...	privat	Angebot	999	control	kleinwagen	1999	manuell	75	2_reihe	150000	
104233	2016-03-26 20:58:26	Citroën_C4_Picasso_2.0_HDI_FAP_EGS6_Exclusive	privat	Angebot	9500	control	bus	2008	automatik	136	c4	125000	
81172	2016-04-01 22:53:21	Volkswagen_Passat_Variant_1.9_TDI_DPF_Comfortline	privat	Angebot	6666	test	kombi	2009	manuell	105	passat	150000	
362697	2016-03-09 14:37:44	BMW_E36_Limo	privat	Angebot	2900	test	NaN	2017	NaN	0	andere	150000	
147593	2016-03-21 08:54:07	Ford_Mondeo_an_Bastler	privat	Angebot	250	control	kombi	1999	manuell	0	mondeo	150000	
254916	2016-03-26 12:45:47	Golf_VII_2.0TDI_DSG_Cup	privat	Angebot	22500	control	limousine	2014	automatik	150	golf	40000	
264392	2016-03-27 16:59:13	Peugeot_307_Premium_4Tuerig_Diesel	privat	Angebot	2790	test	NaN	2017	manuell	109	3_reihe	150000	
	2016-03-19												



Understanding and analyzing the dataset by Correlation:

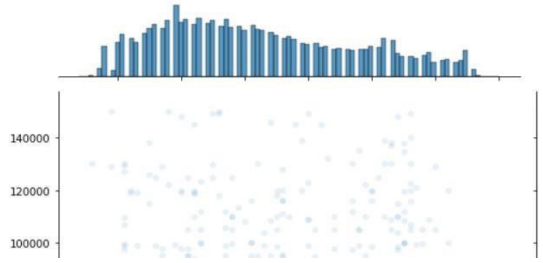
Clearing the null values:

```
[ ] dedups['notRepairedDamage'].fillna(value='not-declared', inplace=True)
    dedups['fuelType'].fillna(value='not-declared', inplace=True)
    dedups['gearbox'].fillna(value='not-declared', inplace=True)
    dedups['vehicleType'].fillna(value='not-declared', inplace=True)
    dedups['model'].fillna(value='not-declared', inplace=True)
```

```
dedups['namelen'] = [min(70, len(n)) for n in dedups['name']]

ax = sns.jointplot(x='namelen',
                   y='price',
                   data=dedups[['namelen','price']],
#                   data=dedups[['namelen','price']][dedups['model']=='golf'],
                   alpha=0.1,
                   size=8)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/axisgrid.py:2182: UserWarning: The `size` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



Preprocessing the Categorical values:

```
labels = ['name', 'gearbox', 'notRepairedDamage', 'model', 'brand', 'fuelType', 'vehicleType']
les = {}

for l in Loading...
    les[l] = preprocessing.LabelEncoder()
    les[l].fit(dedups[l])
    tr = les[l].transform(dedups[l])
    dedups.loc[:, l + '_feat'] = pd.Series(tr, index=dedups.index)

labeled = dedups[ ['price'
                   ,'yearOfRegistration'
                   ,'powerPS'
                   ,'kilometer'
                   ,'monthOfRegistration'
                   , 'namelen']
                 + [x+"_feat" for x in labels]]
```