In [47]:

```
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn import metrics
import statsmodels.api as sm
from sklearn.preprocessing import LabelEncoder,OneHotEncoder
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import RFE
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import statsmodels.api as sfa
from scipy import stats
```

In [48]:

```
df1 = pd.read_csv(r"C:\Users\SOURAV CH\Desktop\data science\projects\dataset.csv",';')
pd.set_option('display.max_columns', None) # inorder to axcess all the columns we need to set the max columns to none
```

In [10]:

```
df1.head(15)
```

Out[10]:

	Nr	region	subregion	state	vot19_14	turnout14	turnout19	turnout19_14	state_abbrev	debt_2013	debt_2014	d
0	1001	Flensburg, Stadt	1.0	Schleswig-Holstein	-0.003027	0.357400	0.562920	0.205520	KS	16.41	16.40	
1	1002	Kiel, Landeshauptstadt	1.0	Schleswig-Holstein	0.060316	0.402589	0.588603	0.186015	KS	12.04	12.03	
2	1003	Lübeck, Hansestadt	1.0	Schleswig-Holstein	0.551817	0.376398	0.546124	0.169726	KS	15.25	15.59	
3	1004	Neumünster, Stadt	1.0	Schleswig-Holstein	2.392481	0.453649	0.482205	0.028556	KS	16.61	16.94	
4	1051	Dithmarschen	1.0	Schleswig-Holstein	3.374651	0.397193	0.544108	0.146915	K	12.52	12.80	
5	1053	Herzogtum Lauenburg	1.0	Schleswig-Holstein	1.646033	0.463842	0.601961	0.138119	K	10.16	10.29	
6	1054	Nordfriesland	1.0	Schleswig-Holstein	-0.013827	0.411905	0.588692	0.176788	K	10.09	10.32	
7	1055	Ostholstein	1.0	Schleswig-Holstein	0.252615	0.424788	0.584264	0.159476	K	10.63	10.80	
8	1056	Pinneberg	1.0	Schleswig-Holstein	0.703336	0.457528	0.629707	0.172180	K	9.67	9.64	
9	1057	Plön	1.0	Schleswig-Holstein	0.307204	0.469130	0.636865	0.167735	K	8.67	8.88	
10	1058	Rendsburg-Eckernförde	1.0	Schleswig-Holstein	0.028835	0.459591	0.624910	0.165319	K	9.16	9.17	
11	1059	Schleswig-Flensburg	1.0	Schleswig-Holstein	-0.508833	0.421130	0.598250	0.177120	K	10.45	10.67	
12	1060	Segeberg	1.0	Schleswig-Holstein	0.948635	0.423453	0.595325	0.171872	K	10.11	10.08	
13	1061	Steinburg	1.0	Schleswig-Holstein	1.952564	0.420752	0.577123	0.156371	K	11.43	11.66	
14	1062	Stormarn	1.0	Schleswig-Holstein	0.523026	0.503786	0.654266	0.150481	K	7.99	7.93	

Simply by looking at the data we can have a first impression of having a high Collinearity as the no of columns are more ,, but we need to further investigate that.

In [4]:

```
print(df1.keys()) # to check the orientation of column names
```

```
Index(['Nr', 'region', 'subregion', 'state', 'vot19_14', 'turnout14',
       'turnout19', 'turnout19_14', 'state_abbrev', 'debt_2013',
       ...
       'f_crime_2015', 'total_suspects_2014', 'foreign_suspects_2014',
       'f_crime_2014', 'total_suspects_2013', 'foreign_suspects_2013',
       'f_crime_2013', 'total_suspects_2012', 'foreign_suspects_2012',
       'f_crime_2012'],
      dtype='object', length=151)
```

In [5]:

```
print(df1.dtypes)
```

```
Nr                          int64
region                     object
subregion                 float64
state                      object
vot19_14                  float64
                           ...
foreign_suspects_2013     float64
f_crime_2013              float64
total_suspects_2012       float64
foreign_suspects_2012     float64
f_crime_2012              float64
Length: 151, dtype: object
```

In [6]:

```
# simply to check the nan values in the columns
df1.isnull().sum().sort_values(ascending = False)
```

Out[6]:

```
empl_agr_2018_2012                 26
empl_agr_2018                      16
birth_balance_2017_2012             9
empl_manuf_2018                     8
empl_manuf_2018_2012                8
                                   ..
age_75_more_2012                    0
graduates_sec_2012                  0
graduates_without_secondary_2012    0
graduates_lower_secondary_2012      0
Nr                                  0
Length: 151, dtype: int64
```

In [7]:

```
df1.describe(include ="all")
# this statement will generally give the overview of data and the information about their behaviou
rs
```

Out[7]:

	Nr	region	subregion	state	vot19_14	turnout14	turnout19	turnout19_14	state_abbrev	debt_2013	debt_2
count	401.000000	401	401.000000	401	401.000000	401.000000	401.000000	401.000000	401	401.000000	401.000

	Nr	region	subregion	state	vot19_14	turnout14	turnout19	turnout19_14	state_abbrev	debt_2013	debt_2
unique	NaN	401	NaN	16	NaN	NaN	NaN	NaN	5	NaN	
top	NaN	Meißen	NaN	Bayern	NaN	NaN	NaN	NaN	LK	NaN	
freq	NaN	1	NaN	96	NaN	NaN	NaN	NaN	251	NaN	
mean	8305.947631	NaN	7.980050	NaN	4.399567	0.471830	0.605913	0.134083	NaN	9.401471	9.502
std	3762.367821	NaN	3.803893	NaN	5.304025	0.071344	0.048212	0.048155	NaN	2.687735	2.729
min	1001.000000	NaN	1.000000	NaN	-2.899875	0.263504	0.476153	0.028556	NaN	3.710000	3.670
25%	5762.000000	NaN	5.000000	NaN	1.091844	0.420203	0.572889	0.095102	NaN	7.440000	7.570
50%	8235.000000	NaN	8.000000	NaN	2.770999	0.475253	0.606205	0.128428	NaN	9.200000	9.220
75%	9676.000000	NaN	9.000000	NaN	4.692135	0.523841	0.639614	0.176785	NaN	10.790000	10.910
max	16077.000000	NaN	16.000000	NaN	21.333849	0.668604	0.743841	0.243963	NaN	19.840000	20.410

In [8]:

```
for i in range(len(df1.index)) :
        print("Nan in rows ",i,":" ,df1.iloc[i].isnull().any().sum())
```

```
Nan in rows  0 : 1
Nan in rows  1 : 1
Nan in rows  2 : 0
Nan in rows  3 : 0
Nan in rows  4 : 0
Nan in rows  5 : 0
Nan in rows  6 : 0
Nan in rows  7 : 0
Nan in rows  8 : 0
Nan in rows  9 : 0
Nan in rows  10 : 0
Nan in rows  11 : 0
Nan in rows  12 : 0
Nan in rows  13 : 0
Nan in rows  14 : 0
Nan in rows  15 : 0
Nan in rows  16 : 0
Nan in rows  17 : 0
Nan in rows  18 : 1
Nan in rows  19 : 0
Nan in rows  20 : 0
Nan in rows  21 : 0
Nan in rows  22 : 0
Nan in rows  23 : 0
Nan in rows  24 : 0
Nan in rows  25 : 1
Nan in rows  26 : 0
Nan in rows  27 : 0
Nan in rows  28 : 0
Nan in rows  29 : 0
Nan in rows  30 : 0
Nan in rows  31 : 0
Nan in rows  32 : 0
Nan in rows  33 : 0
Nan in rows  34 : 0
Nan in rows  35 : 0
Nan in rows  36 : 0
Nan in rows  37 : 0
Nan in rows  38 : 0
Nan in rows  39 : 0
Nan in rows  40 : 0
Nan in rows  41 : 0
Nan in rows  42 : 0
Nan in rows  43 : 0
Nan in rows  44 : 0
Nan in rows  45 : 0
Nan in rows  46 : 0
Nan in rows  47 : 0
Nan in rows  48 : 0
Nan in rows  49 : 0
Nan in rows  50 : 0
Nan in rows  51 : 0
Nan in rows  52 : 0
Nan in rows  53 : 0
Nan in rows  54 : 0
```

```
Nan in rows  55 : 0
Nan in rows  56 : 0
Nan in rows  57 : 0
Nan in rows  58 : 0
Nan in rows  59 : 1
Nan in rows  60 : 1
Nan in rows  61 : 1
Nan in rows  62 : 1
Nan in rows  63 : 0
Nan in rows  64 : 1
Nan in rows  65 : 0
Nan in rows  66 : 0
Nan in rows  67 : 0
Nan in rows  68 : 0
Nan in rows  69 : 1
Nan in rows  70 : 0
Nan in rows  71 : 0
Nan in rows  72 : 0
Nan in rows  73 : 0
Nan in rows  74 : 0
Nan in rows  75 : 0
Nan in rows  76 : 0
Nan in rows  77 : 0
Nan in rows  78 : 1
Nan in rows  79 : 0
Nan in rows  80 : 0
Nan in rows  81 : 0
Nan in rows  82 : 0
Nan in rows  83 : 0
Nan in rows  84 : 0
Nan in rows  85 : 0
Nan in rows  86 : 0
Nan in rows  87 : 0
Nan in rows  88 : 0
Nan in rows  89 : 0
Nan in rows  90 : 0
Nan in rows  91 : 0
Nan in rows  92 : 1
Nan in rows  93 : 0
Nan in rows  94 : 0
Nan in rows  95 : 0
Nan in rows  96 : 0
Nan in rows  97 : 0
Nan in rows  98 : 0
Nan in rows  99 : 0
Nan in rows  100 : 0
Nan in rows  101 : 0
Nan in rows  102 : 0
Nan in rows  103 : 1
Nan in rows  104 : 0
Nan in rows  105 : 0
Nan in rows  106 : 0
Nan in rows  107 : 0
Nan in rows  108 : 0
Nan in rows  109 : 0
Nan in rows  110 : 0
Nan in rows  111 : 0
Nan in rows  112 : 0
Nan in rows  113 : 0
Nan in rows  114 : 0
Nan in rows  115 : 0
Nan in rows  116 : 1
Nan in rows  117 : 0
Nan in rows  118 : 0
Nan in rows  119 : 0
Nan in rows  120 : 0
Nan in rows  121 : 0
Nan in rows  122 : 0
Nan in rows  123 : 0
Nan in rows  124 : 0
Nan in rows  125 : 1
Nan in rows  126 : 0
Nan in rows  127 : 0
Nan in rows  128 : 0
Nan in rows  129 : 0
Nan in rows  130 : 0
Nan in rows  131 : 1
```

Nan in rows 131 : 1
Nan in rows 132 : 0
Nan in rows 133 : 0
Nan in rows 134 : 0
Nan in rows 135 : 1
Nan in rows 136 : 0
Nan in rows 137 : 0
Nan in rows 138 : 0
Nan in rows 139 : 0
Nan in rows 140 : 0
Nan in rows 141 : 0
Nan in rows 142 : 0
Nan in rows 143 : 0
Nan in rows 144 : 0
Nan in rows 145 : 0
Nan in rows 146 : 0
Nan in rows 147 : 1
Nan in rows 148 : 0
Nan in rows 149 : 0
Nan in rows 150 : 0
Nan in rows 151 : 1
Nan in rows 152 : 0
Nan in rows 153 : 0
Nan in rows 154 : 0
Nan in rows 155 : 0
Nan in rows 156 : 0
Nan in rows 157 : 0
Nan in rows 158 : 1
Nan in rows 159 : 1
Nan in rows 160 : 0
Nan in rows 161 : 0
Nan in rows 162 : 0
Nan in rows 163 : 0
Nan in rows 164 : 1
Nan in rows 165 : 1
Nan in rows 166 : 0
Nan in rows 167 : 0
Nan in rows 168 : 0
Nan in rows 169 : 0
Nan in rows 170 : 0
Nan in rows 171 : 0
Nan in rows 172 : 1
Nan in rows 173 : 0
Nan in rows 174 : 0
Nan in rows 175 : 0
Nan in rows 176 : 0
Nan in rows 177 : 0
Nan in rows 178 : 0
Nan in rows 179 : 0
Nan in rows 180 : 1
Nan in rows 181 : 0
Nan in rows 182 : 0
Nan in rows 183 : 0
Nan in rows 184 : 0
Nan in rows 185 : 0
Nan in rows 186 : 0
Nan in rows 187 : 0
Nan in rows 188 : 0
Nan in rows 189 : 0
Nan in rows 190 : 0
Nan in rows 191 : 0
Nan in rows 192 : 0
Nan in rows 193 : 0
Nan in rows 194 : 0
Nan in rows 195 : 0
Nan in rows 196 : 0
Nan in rows 197 : 0
Nan in rows 198 : 0
Nan in rows 199 : 0
Nan in rows 200 : 0
Nan in rows 201 : 0
Nan in rows 202 : 0
Nan in rows 203 : 0
Nan in rows 204 : 0
Nan in rows 205 : 1
Nan in rows 206 : 0
Nan in rows 207 : 0
Nan in rows 208 : 0

Nan in rows 209 : 0
Nan in rows 210 : 0
Nan in rows 211 : 0
Nan in rows 212 : 0
Nan in rows 213 : 1
Nan in rows 214 : 0
Nan in rows 215 : 0
Nan in rows 216 : 0
Nan in rows 217 : 1
Nan in rows 218 : 0
Nan in rows 219 : 0
Nan in rows 220 : 0
Nan in rows 221 : 0
Nan in rows 222 : 0
Nan in rows 223 : 0
Nan in rows 224 : 0
Nan in rows 225 : 0
Nan in rows 226 : 0
Nan in rows 227 : 1
Nan in rows 228 : 0
Nan in rows 229 : 0
Nan in rows 230 : 0
Nan in rows 231 : 0
Nan in rows 232 : 0
Nan in rows 233 : 0
Nan in rows 234 : 0
Nan in rows 235 : 0
Nan in rows 236 : 0
Nan in rows 237 : 0
Nan in rows 238 : 0
Nan in rows 239 : 0
Nan in rows 240 : 0
Nan in rows 241 : 0
Nan in rows 242 : 0
Nan in rows 243 : 0
Nan in rows 244 : 0
Nan in rows 245 : 0
Nan in rows 246 : 0
Nan in rows 247 : 0
Nan in rows 248 : 0
Nan in rows 249 : 0
Nan in rows 250 : 0
Nan in rows 251 : 0
Nan in rows 252 : 0
Nan in rows 253 : 0
Nan in rows 254 : 0
Nan in rows 255 : 0
Nan in rows 256 : 0
Nan in rows 257 : 0
Nan in rows 258 : 0
Nan in rows 259 : 0
Nan in rows 260 : 0
Nan in rows 261 : 0
Nan in rows 262 : 0
Nan in rows 263 : 0
Nan in rows 264 : 0
Nan in rows 265 : 0
Nan in rows 266 : 0
Nan in rows 267 : 0
Nan in rows 268 : 0
Nan in rows 269 : 0
Nan in rows 270 : 0
Nan in rows 271 : 1
Nan in rows 272 : 0
Nan in rows 273 : 1
Nan in rows 274 : 0
Nan in rows 275 : 0
Nan in rows 276 : 0
Nan in rows 277 : 0
Nan in rows 278 : 0
Nan in rows 279 : 0
Nan in rows 280 : 0
Nan in rows 281 : 0
Nan in rows 282 : 0
Nan in rows 283 : 0
Nan in rows 284 : 1
Nan in rows 285 : 0

Nan in rows 285 : 0
Nan in rows 286 : 0
Nan in rows 287 : 1
Nan in rows 288 : 0
Nan in rows 289 : 0
Nan in rows 290 : 0
Nan in rows 291 : 0
Nan in rows 292 : 1
Nan in rows 293 : 1
Nan in rows 294 : 0
Nan in rows 295 : 1
Nan in rows 296 : 0
Nan in rows 297 : 0
Nan in rows 298 : 0
Nan in rows 299 : 0
Nan in rows 300 : 0
Nan in rows 301 : 0
Nan in rows 302 : 1
Nan in rows 303 : 0
Nan in rows 304 : 0
Nan in rows 305 : 0
Nan in rows 306 : 0
Nan in rows 307 : 0
Nan in rows 308 : 0
Nan in rows 309 : 0
Nan in rows 310 : 0
Nan in rows 311 : 0
Nan in rows 312 : 0
Nan in rows 313 : 0
Nan in rows 314 : 0
Nan in rows 315 : 0
Nan in rows 316 : 0
Nan in rows 317 : 0
Nan in rows 318 : 1
Nan in rows 319 : 1
Nan in rows 320 : 1
Nan in rows 321 : 1
Nan in rows 322 : 1
Nan in rows 323 : 1
Nan in rows 324 : 0
Nan in rows 325 : 1
Nan in rows 326 : 1
Nan in rows 327 : 1
Nan in rows 328 : 0
Nan in rows 329 : 0
Nan in rows 330 : 0
Nan in rows 331 : 0
Nan in rows 332 : 0
Nan in rows 333 : 0
Nan in rows 334 : 0
Nan in rows 335 : 1
Nan in rows 336 : 0
Nan in rows 337 : 0
Nan in rows 338 : 0
Nan in rows 339 : 0
Nan in rows 340 : 1
Nan in rows 341 : 0
Nan in rows 342 : 0
Nan in rows 343 : 0
Nan in rows 344 : 0
Nan in rows 345 : 0
Nan in rows 346 : 0
Nan in rows 347 : 0
Nan in rows 348 : 0
Nan in rows 349 : 0
Nan in rows 350 : 0
Nan in rows 351 : 0
Nan in rows 352 : 0
Nan in rows 353 : 0
Nan in rows 354 : 0
Nan in rows 355 : 0
Nan in rows 356 : 0
Nan in rows 357 : 0
Nan in rows 358 : 0
Nan in rows 359 : 0
Nan in rows 360 : 0
Nan in rows 361 : 0
Nan in rows 362 : 0

```
Nan in rows  363 : 0
Nan in rows  364 : 1
Nan in rows  365 : 0
Nan in rows  366 : 0
Nan in rows  367 : 0
Nan in rows  368 : 0
Nan in rows  369 : 0
Nan in rows  370 : 0
Nan in rows  371 : 0
Nan in rows  372 : 0
Nan in rows  373 : 0
Nan in rows  374 : 0
Nan in rows  375 : 0
Nan in rows  376 : 0
Nan in rows  377 : 0
Nan in rows  378 : 0
Nan in rows  379 : 0
Nan in rows  380 : 1
Nan in rows  381 : 1
Nan in rows  382 : 1
Nan in rows  383 : 0
Nan in rows  384 : 0
Nan in rows  385 : 0
Nan in rows  386 : 0
Nan in rows  387 : 0
Nan in rows  388 : 0
Nan in rows  389 : 0
Nan in rows  390 : 0
Nan in rows  391 : 0
Nan in rows  392 : 0
Nan in rows  393 : 0
Nan in rows  394 : 0
Nan in rows  395 : 1
Nan in rows  396 : 0
Nan in rows  397 : 0
Nan in rows  398 : 0
Nan in rows  399 : 0
Nan in rows  400 : 0
```

In [9]:

```
# accessing the row and column having the null value
df1.iloc[17:19] # then by using the row and column retreving functions we can access the columns h
aving the nan values
```

Out[9]:

	Nr	region	subregion	state	vot19_14	turnout14	turnout19	turnout19_14	state_abbrev	debt_2013	debt_2014	deb
17	3102	Salzgitter, Stadt	3.0	Niedersachsen	9.198355	0.459825	0.511638	0.051813	KS	12.16	12.51	
18	3103	Wolfsburg, Stadt	3.0	Niedersachsen	3.600941	0.412644	0.564967	0.152323	KS	8.17	7.99	

From the above describe function we can see that the mean and median have almost approximately the same values .

- **But as there is a huge difference between then min and max values our data might contain Outliers or some influential points.**
- **How do we do that , lets get to know .**

In [10]:

```
# simply to check the nan values in the columns
df1.isnull().sum().sort_values(ascending = False)
```

Out[10]:

```
ampl_sav_2010_2012                  26
```

```
empl_agr_2018_2012                  26
empl_agr_2018                       16
birth_balance_2017_2012              9
empl_manuf_2018                      8
empl_manuf_2018_2012                 8
                                    ..
age_75_more_2012                     0
graduates_sec_2012                   0
graduates_without_secondary_2012     0
graduates_lower_secondary_2012       0
Nr                                   0
Length: 151, dtype: int64
```

In [11]:

```
# this line will just print the rows having nan values in the colums based on the total nan value count in it(i.e Ascending order)
df1.assign(Count_NA = lambda x: x.isnull().sum(axis=1)).sort_values('Count_NA', ascending=False)
```

Out[11]:

	Nr	region	subregion	state	vot19_14	turnout14	turnout19	turnout19_14	state_abbrev	debt_2013	debt_20
60	3462	Wittmund	3.0	Niedersachsen	3.795097	0.477998	0.529910	0.051912	LK	10.51	10
381	16054	Suhl, Stadt	16.0	Thüringen	10.577550	0.451281	0.564866	0.113586	KS	10.03	10
61	4011	Bremen, Stadt	4.0	Bremen	1.393191	0.415112	0.652452	0.237340	KS	12.67	12
321	10044	Saarlouis	10.0	Saarland	3.039576	0.542929	0.674194	0.131264	LK	9.81	9
380	16053	Jena, Stadt	16.0	Thüringen	5.929222	0.522503	0.649774	0.127272	KS	5.81	5
...	...	...	...	...	...	...	...	...	...	...	...
138	6633	Kassel	6.0	Hessen	0.131689	0.413184	0.565107	0.151923	LK	9.06	9
137	6632	Hersfeld-Rotenburg	6.0	Hessen	3.386571	0.389272	0.551690	0.162418	LK	9.11	9
136	6631	Fulda	6.0	Hessen	3.861435	0.391362	0.571427	0.180065	LK	7.52	7
134	6535	Vogelsbergkreis	6.0	Hessen	2.893113	0.417771	0.563673	0.145902	LK	8.41	8
400	16077	Altenburger Land	16.0	Thüringen	19.110053	0.483017	0.565661	0.082643	LK	8.29	8

401 rows × 152 columns

- From the describe function we saw mean,median were approximately same , but as we suspected ouliers we will visualize them ,But in order to visualize them uisng box plot we remove nan values from data .
- For nan value **IMPUTATION** use **Median** as there are possible ouliers

In [49]:

```
df1.fillna(df1.median(),inplace = True)
```

In [12]:

```
df1.isnull().all().sum()
# wehave removed all the nan values from the data now lets visualize the attributes
```

Out[12]:

```
0
```

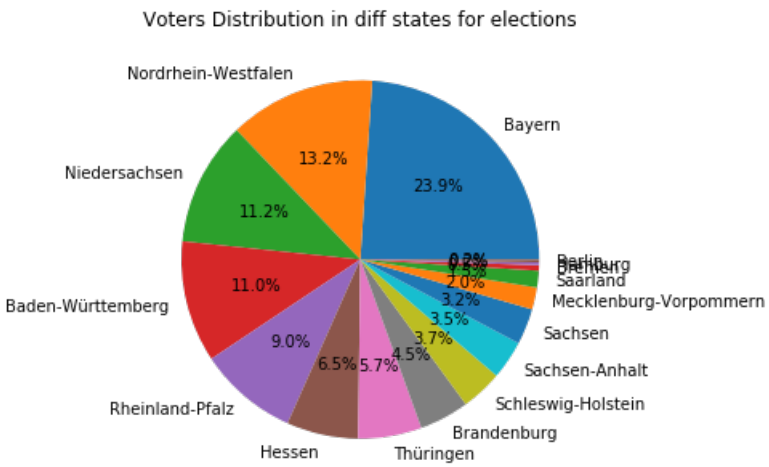
Data visualization:

- Brief knowledge of the data.
- Indetifying the target varaible

- identiying the turnout ratio patterns regarding different attributes.
- Identifying the key attributes that constitute to the target variable.
- uni,Bi, Multivariate analysis.
- knowledge of the distribution inside the data.

In [15]:

```
fig = plt.figure(figsize=(5,10))
df1['state'].value_counts().plot(kind = 'pie', autopct='%.1f%%')
plt.ylabel(" ", fontsize = 25)
plt.title("Voters Distribution in diff states for elections")
print("")
```



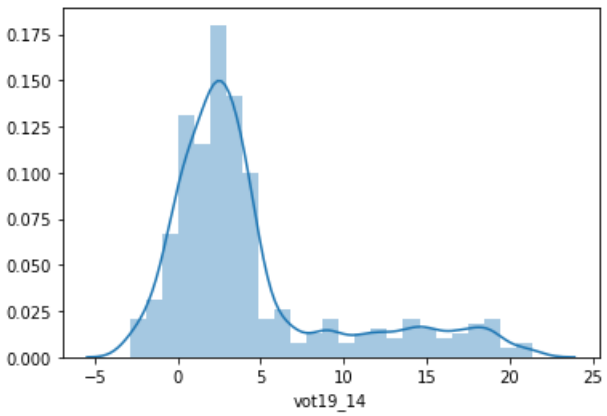
Bayern state people have voted highest for AFD in the election (23.9%) . i.e change rates are higher for this state

- **Berlin state is the state voted the least for AFD (0.2%).**

In [125]:

```
sns.distplot(df1['vot19_14']);
#skewness and kurtosis
print("Skewness: %f" % df1['vot19_14'].skew())
print("Kurtosis: %f" % df1['vot19_14'].kurt())
```

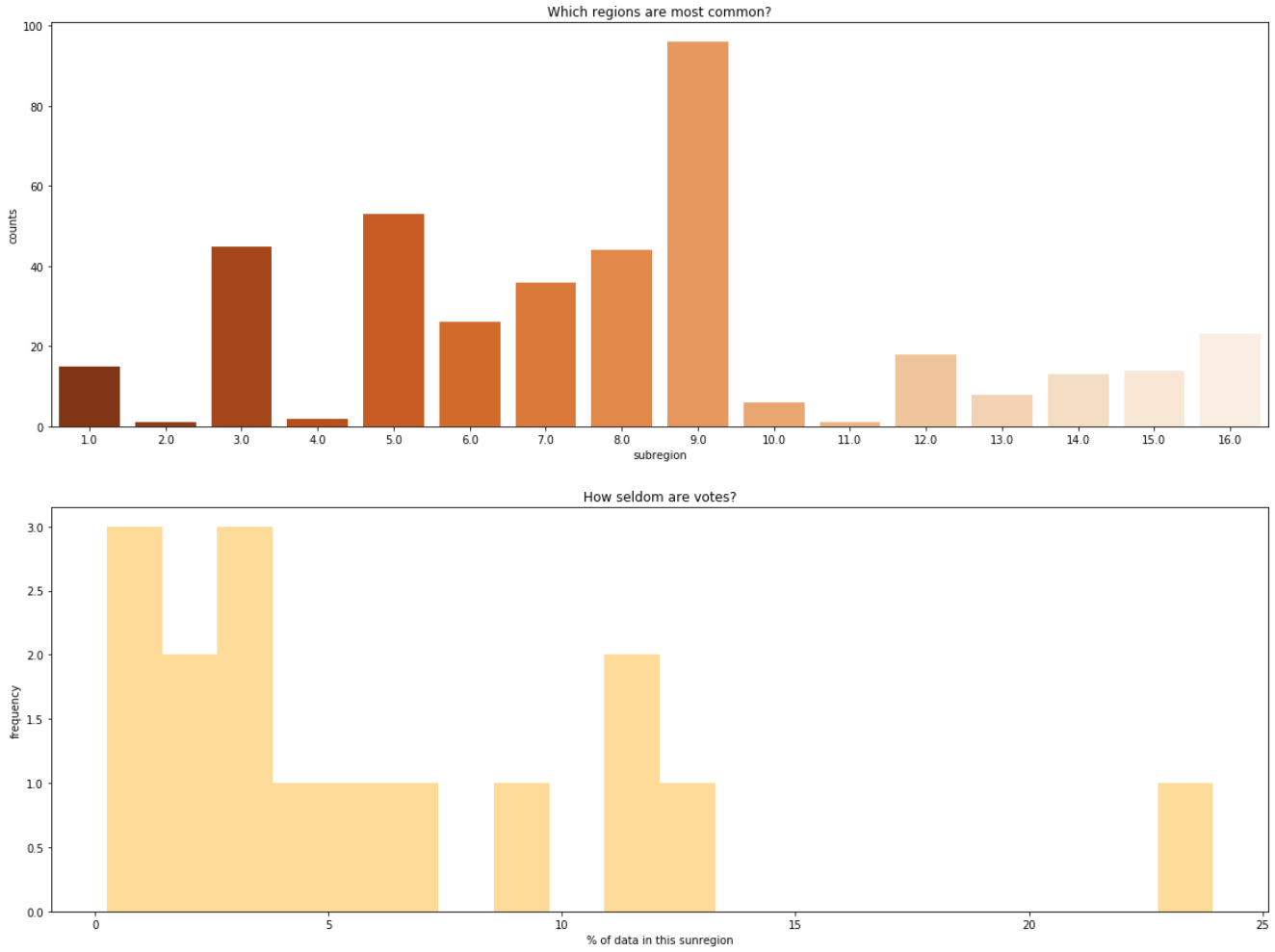
```
Skewness: 1.511796
Kurtosis: 1.474949
```



- From the above plot we can say that the target column is **right** skewed as we can clearly see that mean > median. so most of the data point fall towars the positive skew.
- Which is why we have imputed the missing values with median.
- the data is platykurtic less than the meso or normal distribution.

In [70]:

```
vote_counts = df1.subregion.value_counts().sort_values(ascending=False)
fig, ax = plt.subplots(2,1,figsize=(20,15))
sns.barplot(vote_counts.iloc[0:20].index,
            vote_counts.iloc[0:20].values,
            ax = ax[0], palette="Oranges_r")
ax[0].set_ylabel("counts")
ax[0].set_xlabel("subregion")
ax[0].set_title("Which regions are most common?");
sns.distplot(np.round(vote_counts/df1.shape[0]*100,2),
             kde=False,
             bins=20,
             ax=ax[1], color="Orange")
ax[1].set_title("How seldom are votes?")
ax[1].set_xlabel("% of data in this sunregion")
ax[1].set_ylabel("frequency");
```



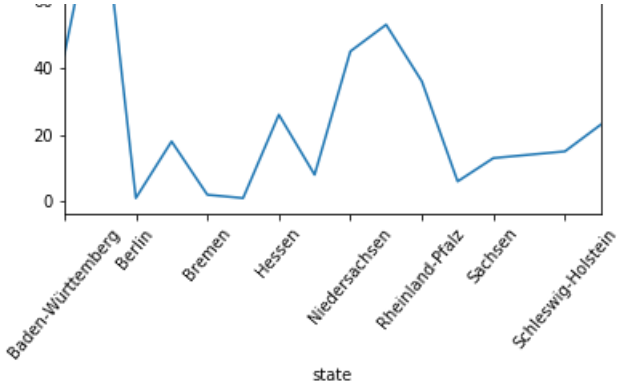
The above plot gives us the information about the percentage of voters in 19_14 election change in different subregions

In [10]:

```
n_by_state = df1.groupby("state")["turnout19"].count().plot()
plt.xticks(rotation=50)
```

Out[10]:

```
(array([ 0.,  2.,  4.,  6.,  8., 10., 12., 14., 16.]),
 <a list of 9 Text xticklabel objects>)
```

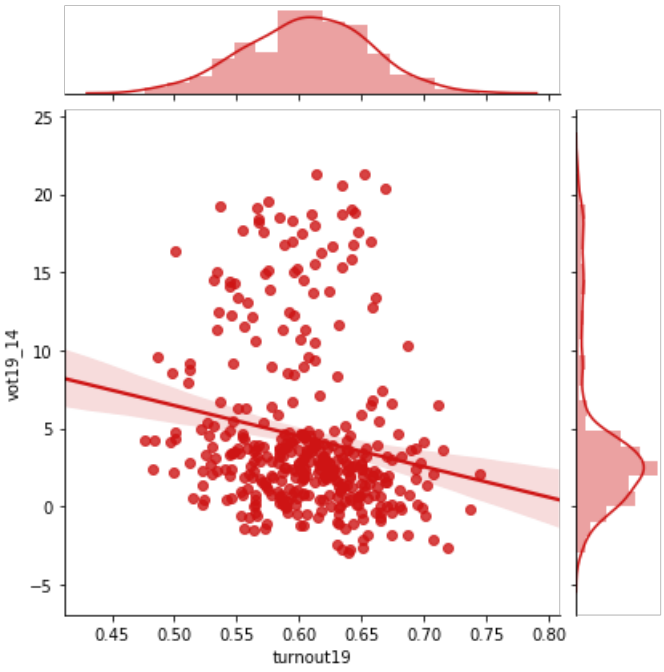


In [13]:

```
sns.jointplot(df1.loc[:,'turnout19'], df1.loc[:,'vot19_14'], kind="regg", color="#ce1414")
```

Out[13]:

```
<seaborn.axisgrid.JointGrid at 0x26671092648>
```

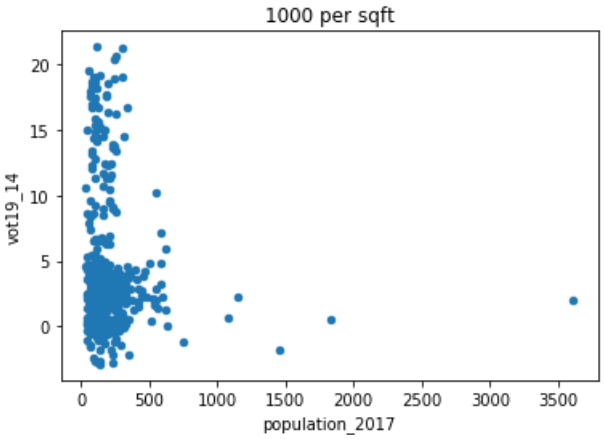


In [12]:

```
df1.plot.scatter(x='population_2017', y='vot19_14', title='1000 per sqft')
```

Out[12]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x26670058e08>
```

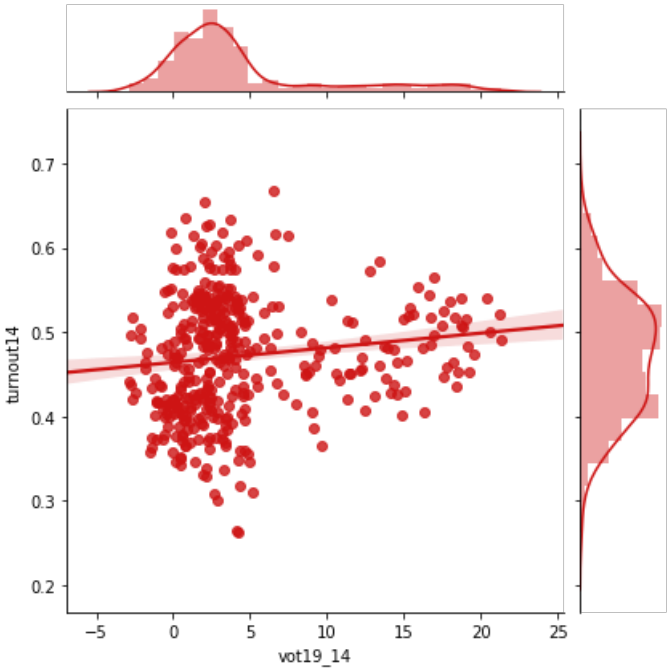


In [82]:

```
sns.jointplot(df1.loc[:,'vot19_14'], df1.loc[:,'turnout14'], kind="regg", color="#ce1414")
```

Out[82]:

```
<seaborn.axisgrid.JointGrid at 0x1f4ee56dbe0>
```



In [13]:

```
df1["vot19_14_lod"] = np.log(df1["vot19_14"])
```

```
C:\Users\SOURAV CH\Anaconda3\lib\site-packages\pandas\core\series.py:853: RuntimeWarning: invalid
value encountered in log
  result = getattr(ufunc, method)(*inputs, **kwargs)
```

In [15]:

```
df1["vot19_14_lod"]
```

Out[15]:

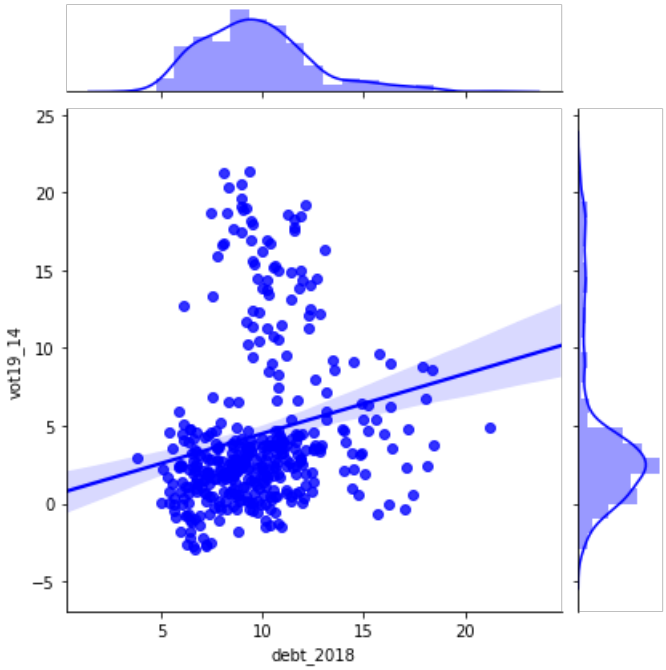
```
0           NaN
1     -2.808166
2     -0.594539
3      0.872331
4      1.216292
         ...
396    2.929166
397    2.594478
398    2.929572
399    2.765156
400    2.950215
Name: vot19_14_lod, Length: 401, dtype: float64
```

In [81]:

```
sns.jointplot(df1.loc[:,'debt_2018'], df1.loc[:,'vot19_14'], kind="regg", color="blue")
```

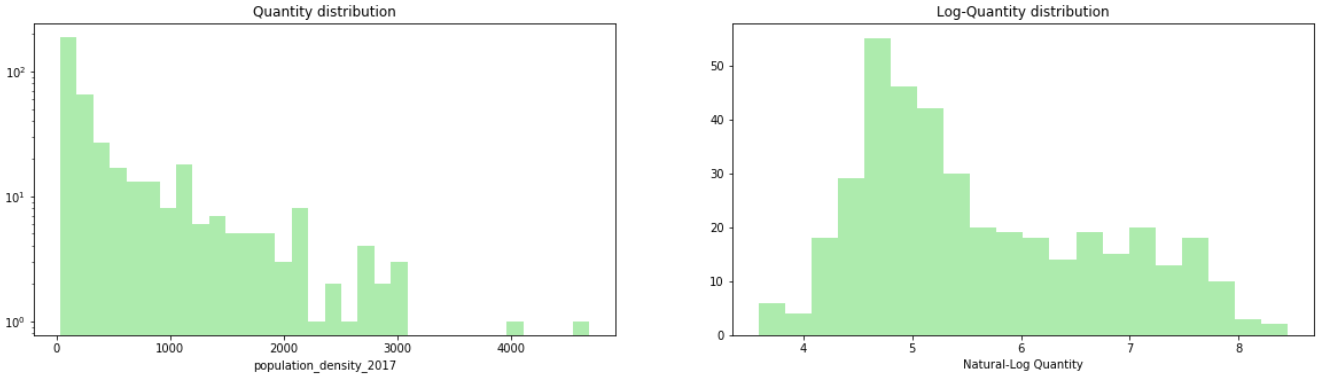
Out[81]:

```
<seaborn.axisgrid.JointGrid at 0x1f4ee306358>
```



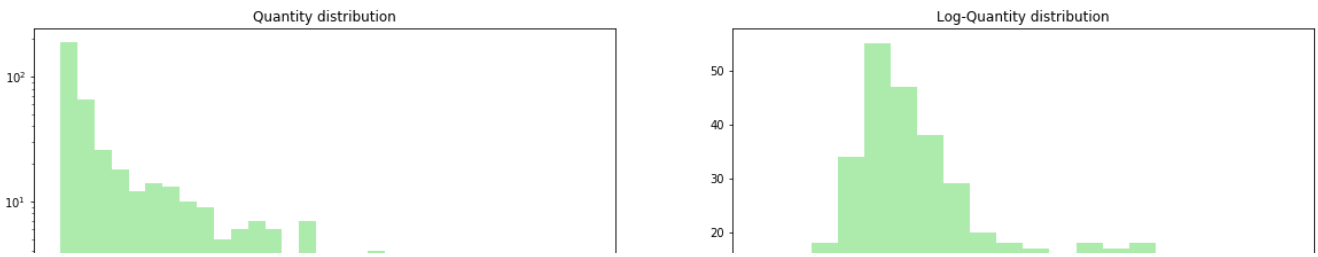
In [86]:

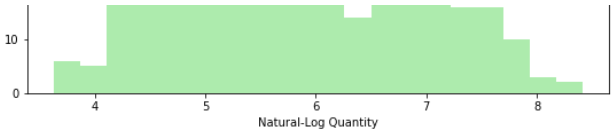
```
fig, ax = plt.subplots(1,2,figsize=(20,5))
sns.distplot(df1.population_density_2017, ax=ax[0], kde=False, color="limegreen");
sns.distplot(np.log(df1.population_density_2017), ax=ax[1], bins=20, kde=False, color="limegreen");
ax[0].set_title("Quantity distribution")
ax[0].set_yscale("log")
ax[1].set_title("Log-Quantity distribution")
ax[1].set_xlabel("Natural-Log Quantity");

# below plot is to visualize the population density of all the sates collectively
```



In [17]:

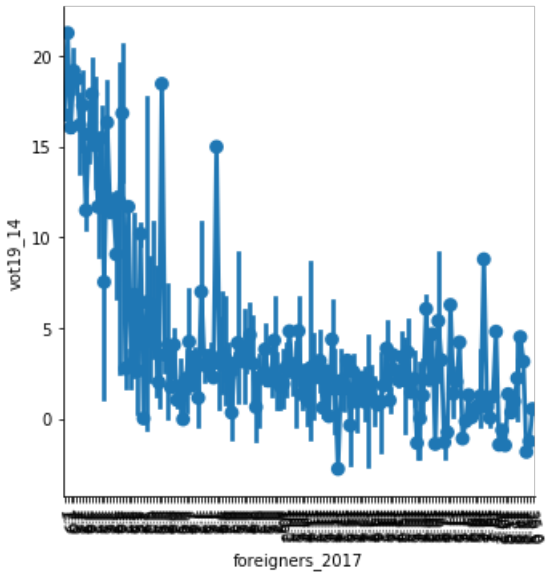
```
fig, ax = plt.subplots(1,2,figsize=(20,5))
sns.distplot(df1.population_density_2012, ax=ax[0], kde=False, color="limegreen");
sns.distplot(np.log(df1.population_density_2012), ax=ax[1], bins=20, kde=False, color="limegreen");
ax[0].set_title("Quantity distribution")
ax[0].set_yscale("log")
ax[1].set_title("Log-Quantity distribution")
ax[1].set_xlabel("Natural-Log Quantity");

# below plot is to visualize the population density of all the sates collectively
```





In [114]:

```
sns.factorplot('foreigners_2017','vot19_14',data=df1,samples = 10)
plt.xticks(rotation=1000000000)
plt.show()
```

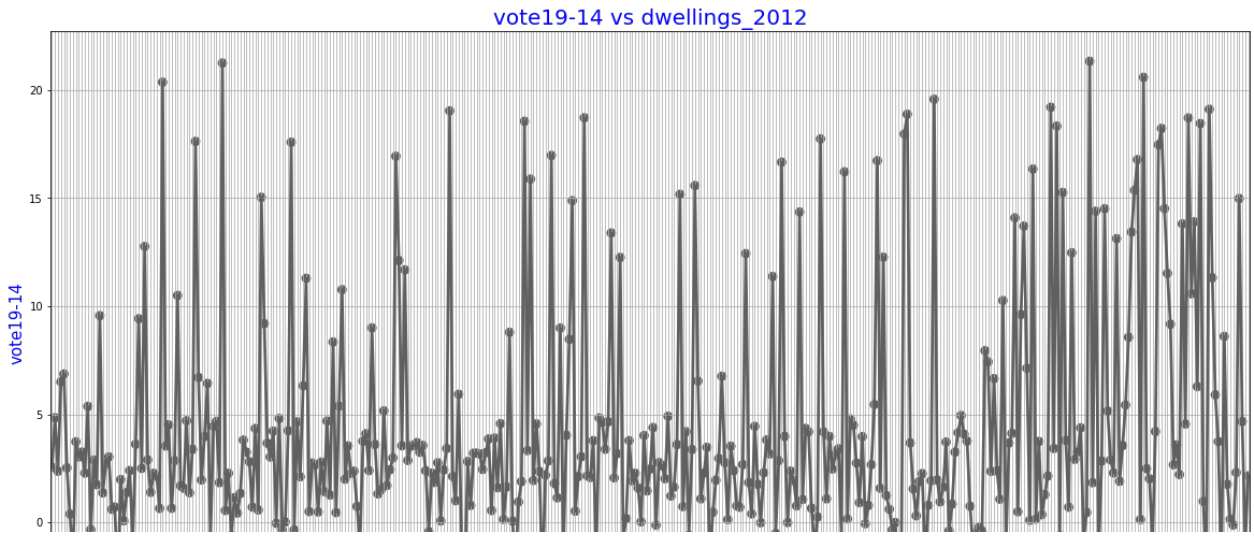


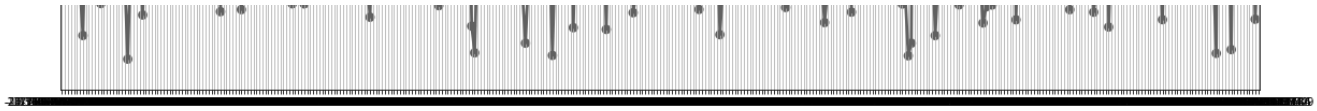
In [ ]:

```
fig, ax = plt.subplots(1,2,figsize=(20,5))
sns.distplot(daily_data.Quantity.values, kde=True, ax=ax[0], color="Orange", bins=30);
sns.distplot(np.log(daily_data.Quantity.values), kde=True, ax=ax[1], color="Orange", bins=30);
ax[0].set_xlabel("Number of daily product sales");
ax[0].set_ylabel("Frequency");
ax[0].set_title("How many products are sold per day?");
```

In [99]:

```
f,ax1 = plt.subplots(figsize =(20,10))
sns.pointplot(x='dwellings_2017_2012',y='vot19_14',data=df1,color='#606060',alpha=0.8)
plt.xlabel('dwellings_2017_2012',fontsize = 15,color='blue')
plt.ylabel('vote19-14',fontsize = 15,color='blue')
plt.title('vote19-14 vs dwellings_2012',fontsize = 20,color='blue')
plt.grid()
plt.show()
```





dwellings_2017_2012

From the above plot we can infer that as the no of dwellings got increased by 2019 from 2012 , population got increased which might be a major factor for AFd turnout to parliament when compared to 2014

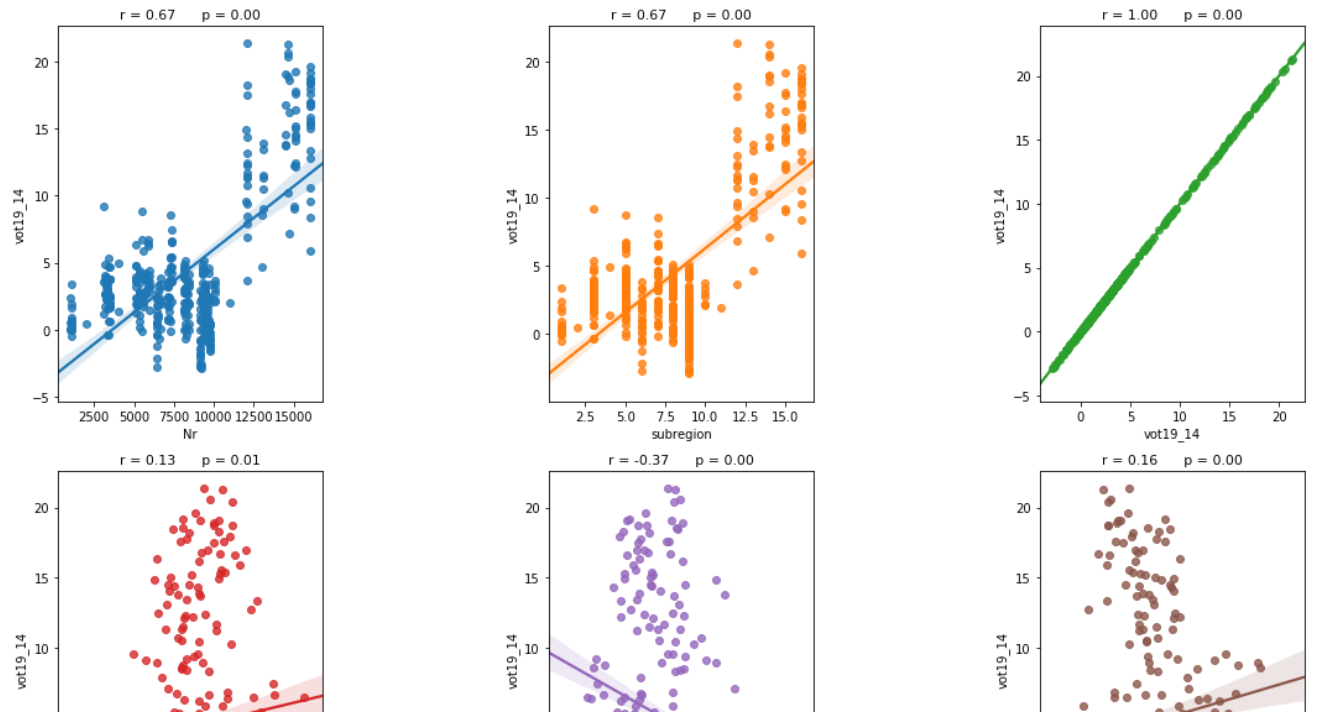
In [15]:

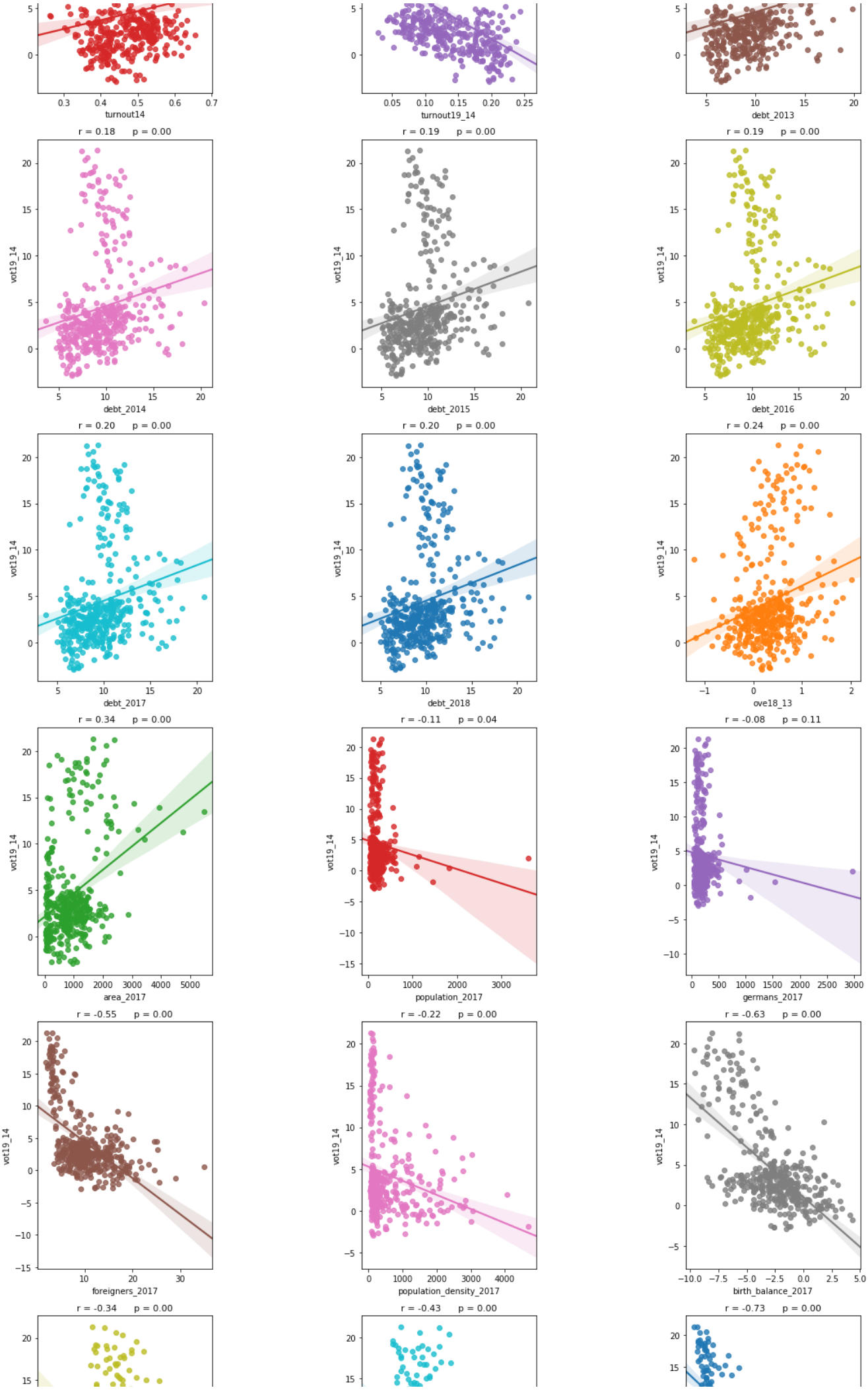
```
numerical_feats = df1.dtypes[df1.dtypes != "object"].index
print("Number of Numerical features: ", len(numerical_feats))

categorical_feats = df1.dtypes[df1.dtypes == "object"].index
print("Number of Categorical features: ", len(categorical_feats))
```

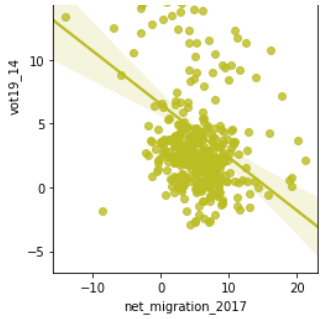
```
Number of Numerical features:  148
Number of Categorical features:  3
```

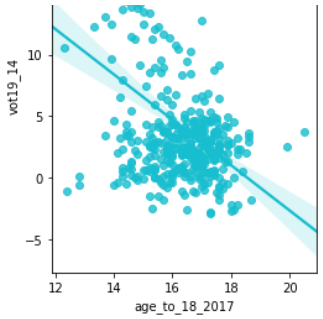
In [152]:

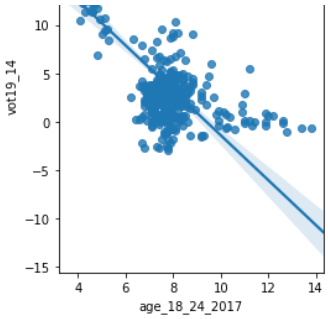
```
target = 'vot19_14'
nr_rows = 55
nr_cols = 3

fig, axs = plt.subplots(nr_rows, nr_cols, figsize=(nr_cols*5,nr_rows*5))

li_num_feats = list(numerical_feats)
li_not_plot = ['state', 'vote19_14', 'turnout19']
li_plot_num_feats = [c for c in list(numerical_feats) if c not in li_not_plot]


for r in range(0,nr_rows):
    for c in range(0,nr_cols):
        i = r*nr_cols+c
        if i < len(li_plot_num_feats):
            sns.regplot(df1[li_plot_num_feats[i]], df1[target], ax = axs[r][c])
            stp = stats.pearsonr(df1[li_plot_num_feats[i]], df1[target])
            #axs[r][c].text(0.4,0.9,"title",fontsize=7)
            str_title = "r = " + "{0:.2f}".format(stp[0]) + "      " "p = " + "{0:.2f}".format(stp[1
])
            axs[r][c].set_title(str_title,fontsize=11)

plt.tight_layout()
plt.show()
```

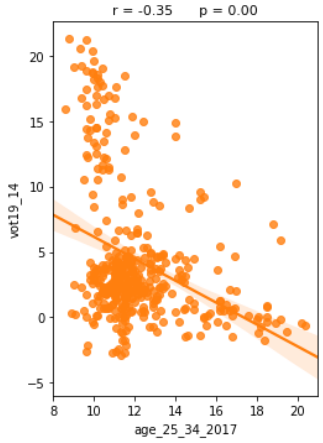


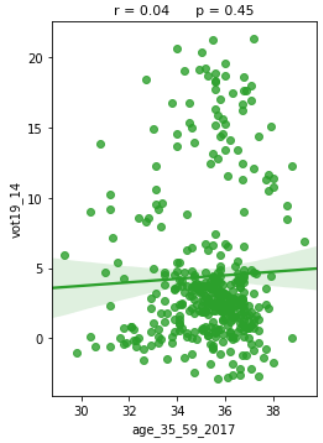

turnout14
turnout19_14
debt_2013
r = 0.18 p = 0.00
r = 0.19 p = 0.00
r = 0.19 p = 0.00
vot19_14
debt_2014
debt_2015
debt_2016
r = 0.20 p = 0.00
r = 0.20 p = 0.00
r = 0.24 p = 0.00
debt_2017
debt_2018
ove18_13
r = 0.34 p = 0.00
r = -0.11 p = 0.04
r = -0.08 p = 0.11
area_2017
population_2017
germans_2017
r = -0.55 p = 0.00
r = -0.22 p = 0.00
r = -0.63 p = 0.00
foreigners_2017
population_density_2017
birth_balance_2017
r = -0.34 p = 0.00
r = -0.43 p = 0.00
r = -0.73 p = 0.00

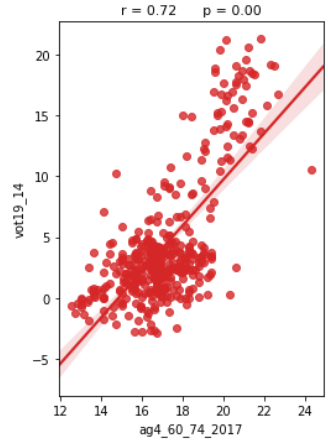

vot19_14
net_migration_2017

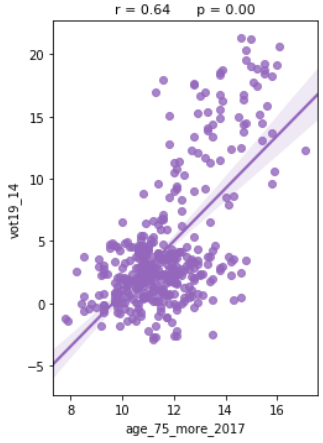

vot19_14
age_to_18_2017

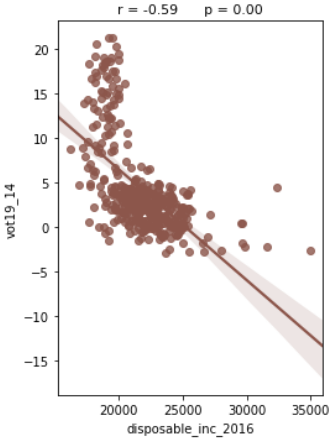

vot19_14
age_18_24_2017

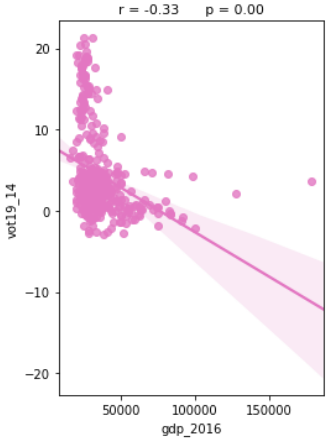

r = -0.35 p = 0.00
vot19_14
age_25_34_2017

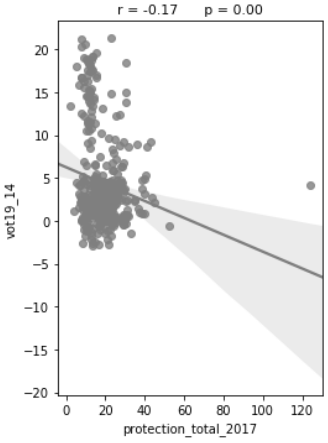

r = 0.04 p = 0.45
vot19_14
age_35_59_2017

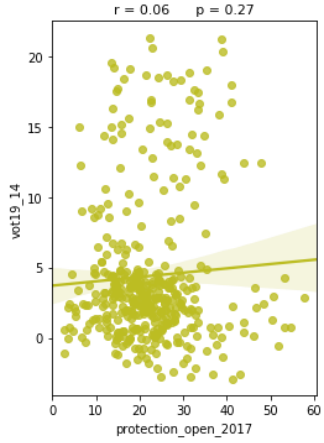

r = 0.72 p = 0.00
vot19_14
ag4_60_74_2017

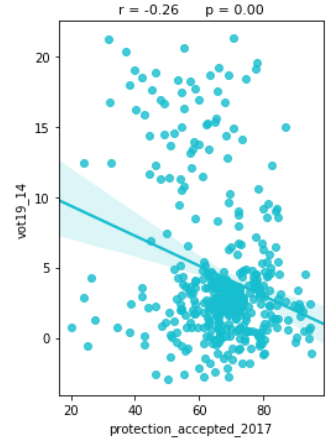

r = 0.64 p = 0.00
vot19_14
age_75_more_2017

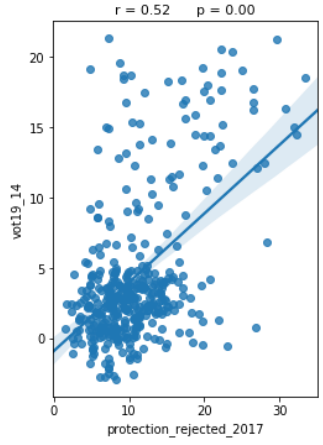

r = -0.59 p = 0.00
vot19_14
disposable_inc_2016

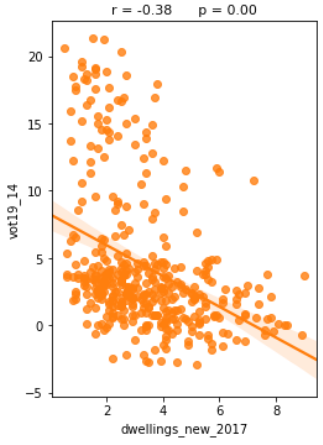

r = -0.33 p = 0.00
vot19_14
gdp_2016

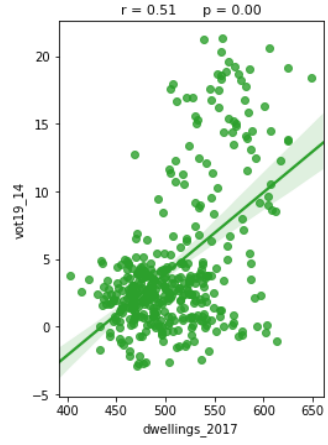

r = -0.17 p = 0.00
vot19_14
protection_total_2017

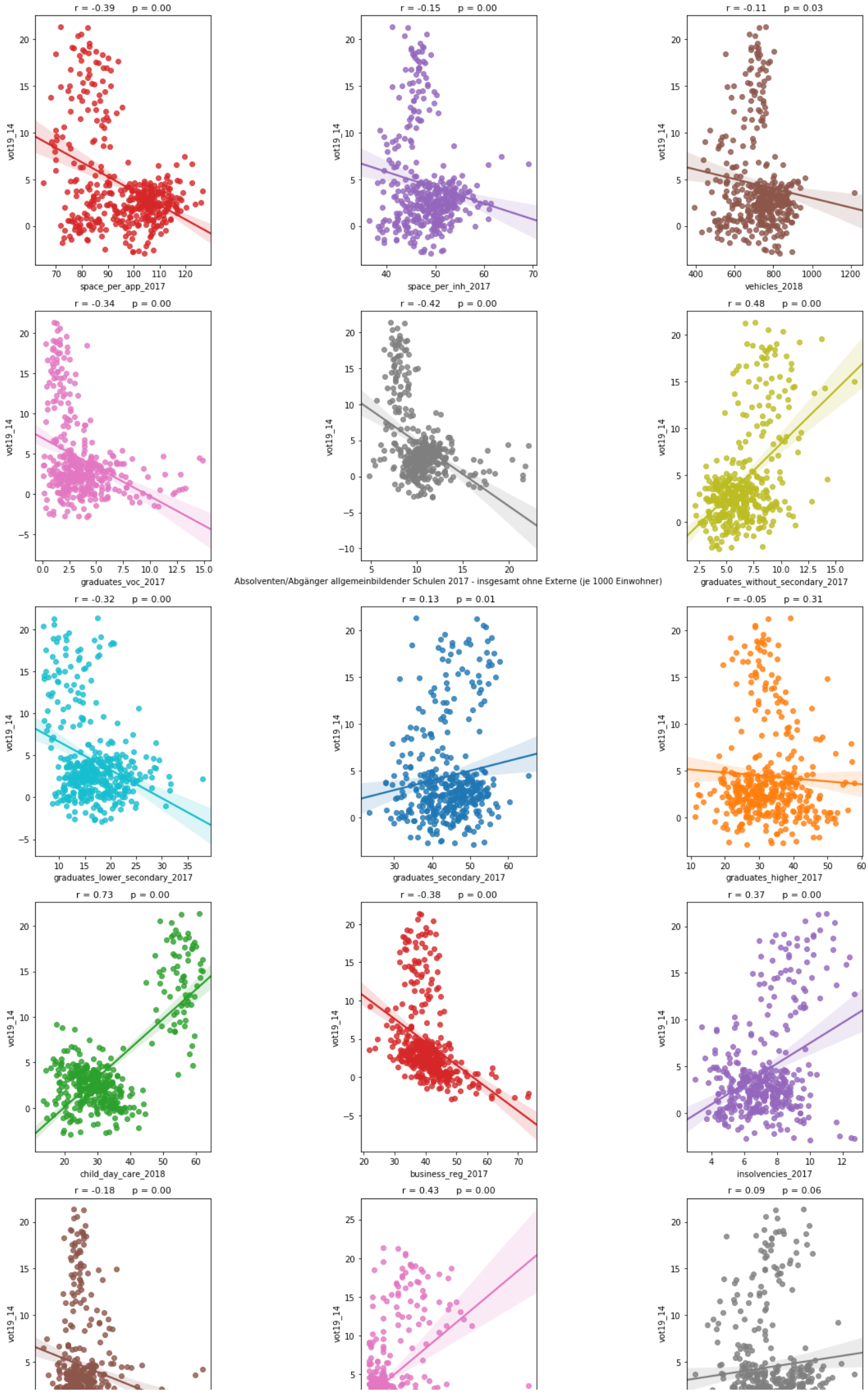

r = 0.06 p = 0.27
vot19_14
protection_open_2017

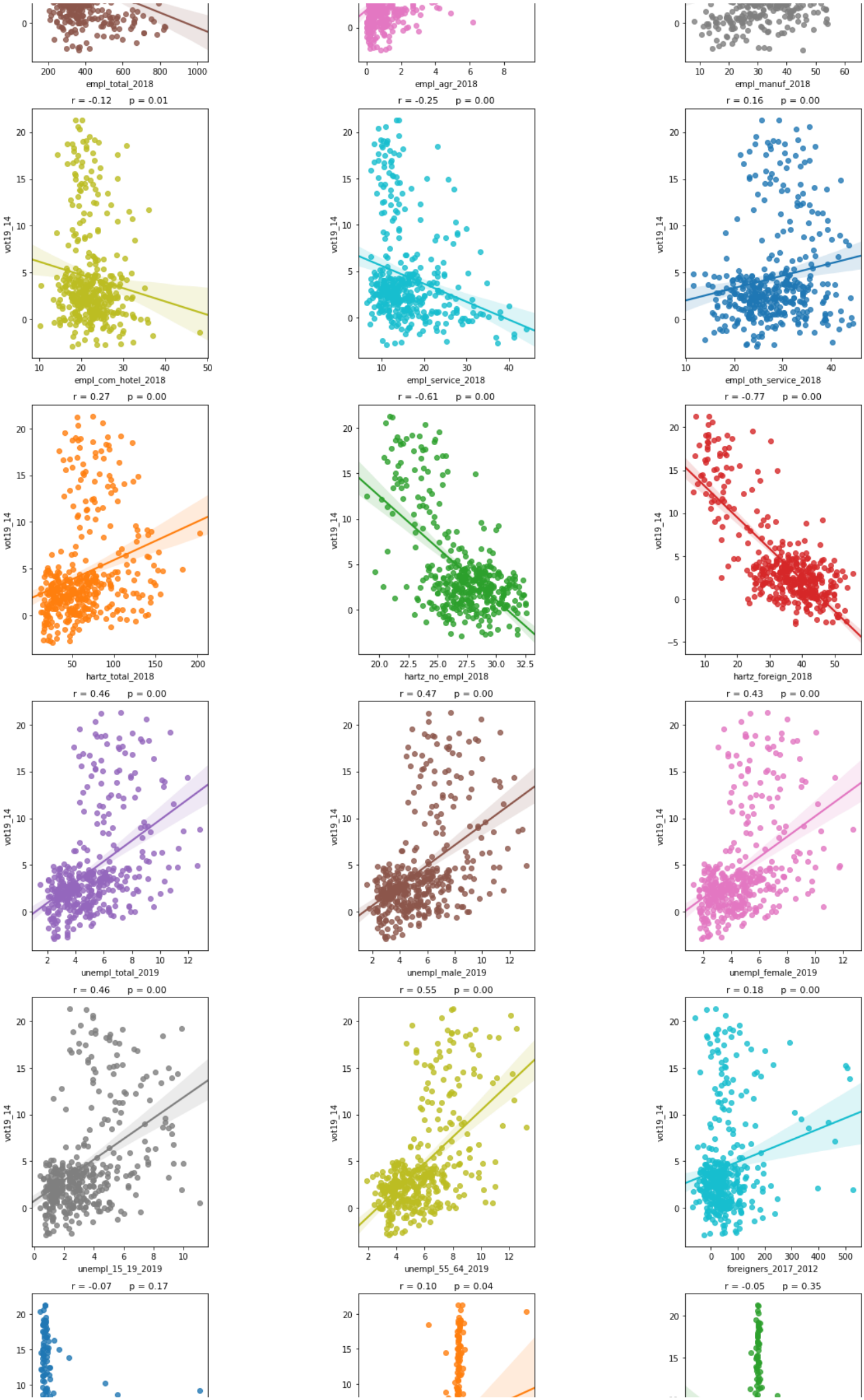

r = -0.26 p = 0.00
vot19_14
protection_accepted_2017

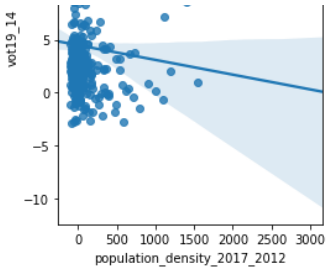

r = 0.52 p = 0.00
vot19_14
protection_rejected_2017

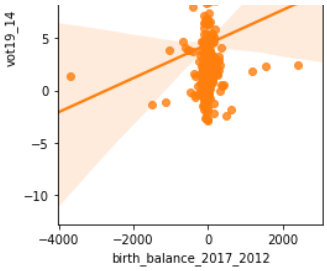

r = -0.38 p = 0.00
vot19_14
dwellings_new_2017

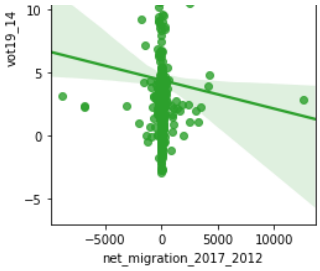

r = 0.51 p = 0.00
vot19_14
dwellings_2017

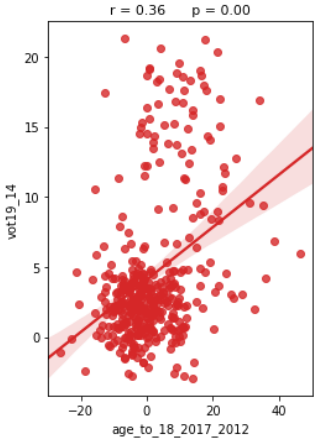

r = -0.39 p = 0.00
vot19_14
space_per_app_2017
r = -0.15 p = 0.00
vot19_14
space_per_inh_2017
r = -0.11 p = 0.03
vot19_14
vehicles_2018
r = -0.34 p = 0.00
vot19_14
graduates_voc_2017
r = -0.42 p = 0.00
vot19_14
Absolventen/Abgänger allgemeinbildender Schulen 2017 - insgesamt ohne Externe (je 1000 Einwohner)
r = 0.48 p = 0.00
vot19_14
graduates_without_secondary_2017
r = -0.32 p = 0.00
vot19_14
graduates_lower_secondary_2017
r = 0.13 p = 0.01
vot19_14
graduates_secondary_2017
r = -0.05 p = 0.31
vot19_14
graduates_higher_2017
r = 0.73 p = 0.00
vot19_14
child_day_care_2018
r = -0.38 p = 0.00
vot19_14
business_reg_2017
r = 0.37 p = 0.00
vot19_14
insolvencies_2017
r = -0.18 p = 0.00
vot19_14
r = 0.43 p = 0.00
vot19_14
r = 0.09 p = 0.06
vot19_14

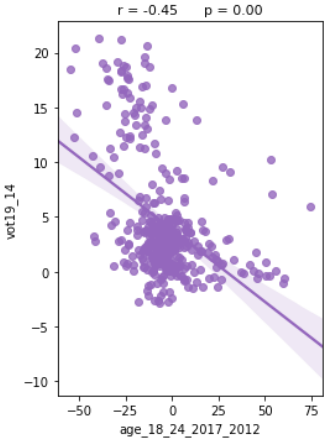

empl_total_2018
empl_agr_2018
empl_manuf_2018
r = -0.12 p = 0.01
r = -0.25 p = 0.00
r = 0.16 p = 0.00
vot19_14
empl_com_hotel_2018
empl_service_2018
empl_oth_service_2018
r = 0.27 p = 0.00
r = -0.61 p = 0.00
r = -0.77 p = 0.00
hartz_total_2018
hartz_no_empl_2018
hartz_foreign_2018
r = 0.46 p = 0.00
r = 0.47 p = 0.00
r = 0.43 p = 0.00
unempl_total_2019
unempl_male_2019
unempl_female_2019
r = 0.46 p = 0.00
r = 0.55 p = 0.00
r = 0.18 p = 0.00
unempl_15_19_2019
unempl_55_64_2019
foreigners_2017_2012
r = -0.07 p = 0.17
r = 0.10 p = 0.04
r = -0.05 p = 0.35

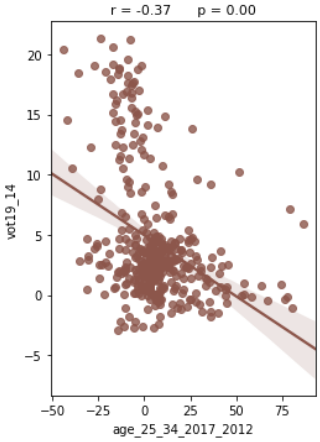

vot19_14
population_density_2017_2012

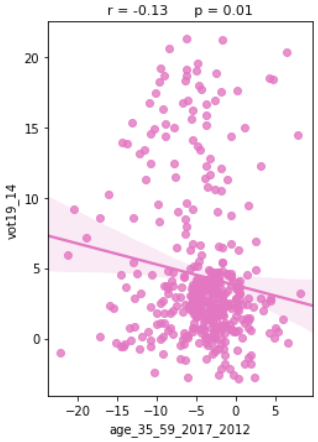

vot19_14
birth_balance_2017_2012

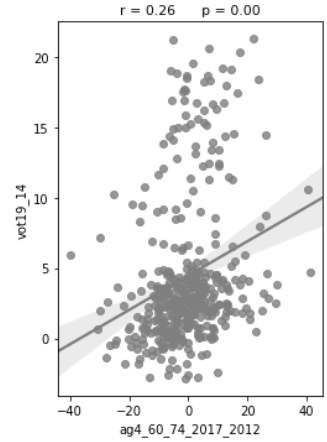

vot19_14
net_migration_2017_2012

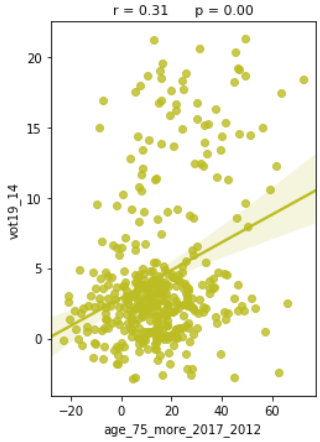

r = 0.36 p = 0.00
vot19_14
age_to_18_2017_2012

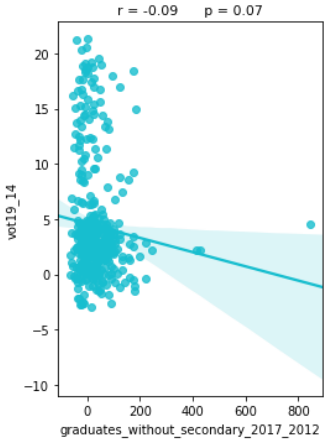

r = -0.45 p = 0.00
vot19_14
age_18_24_2017_2012

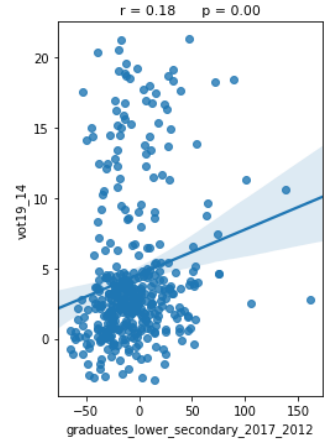

r = -0.37 p = 0.00
vot19_14
age_25_34_2017_2012

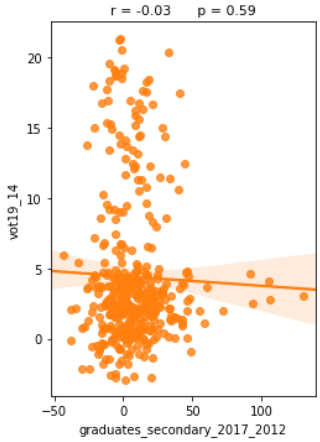

r = -0.13 p = 0.01
vot19_14
age_35_59_2017_2012

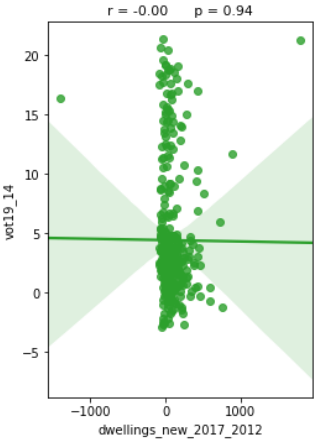

r = 0.26 p = 0.00
vot19_14
ag4_60_74_2017_2012

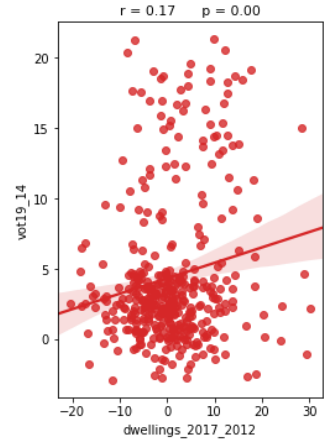

r = 0.31 p = 0.00
vot19_14
age_75_more_2017_2012

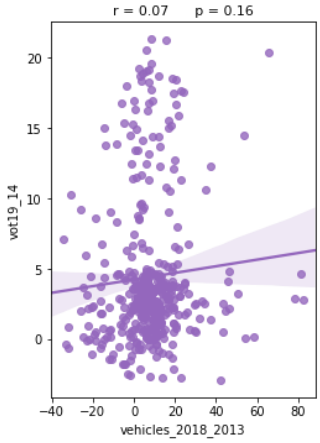

r = -0.09 p = 0.07
vot19_14
graduates_without_secondary_2017_2012

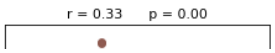

r = 0.18 p = 0.00
vot19_14
graduates_lower_secondary_2017_2012


r = -0.03 p = 0.59
vot19_14
graduates_secondary_2017_2012

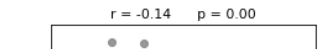

r = -0.00 p = 0.94
vot19_14
dwellings_new_2017_2012

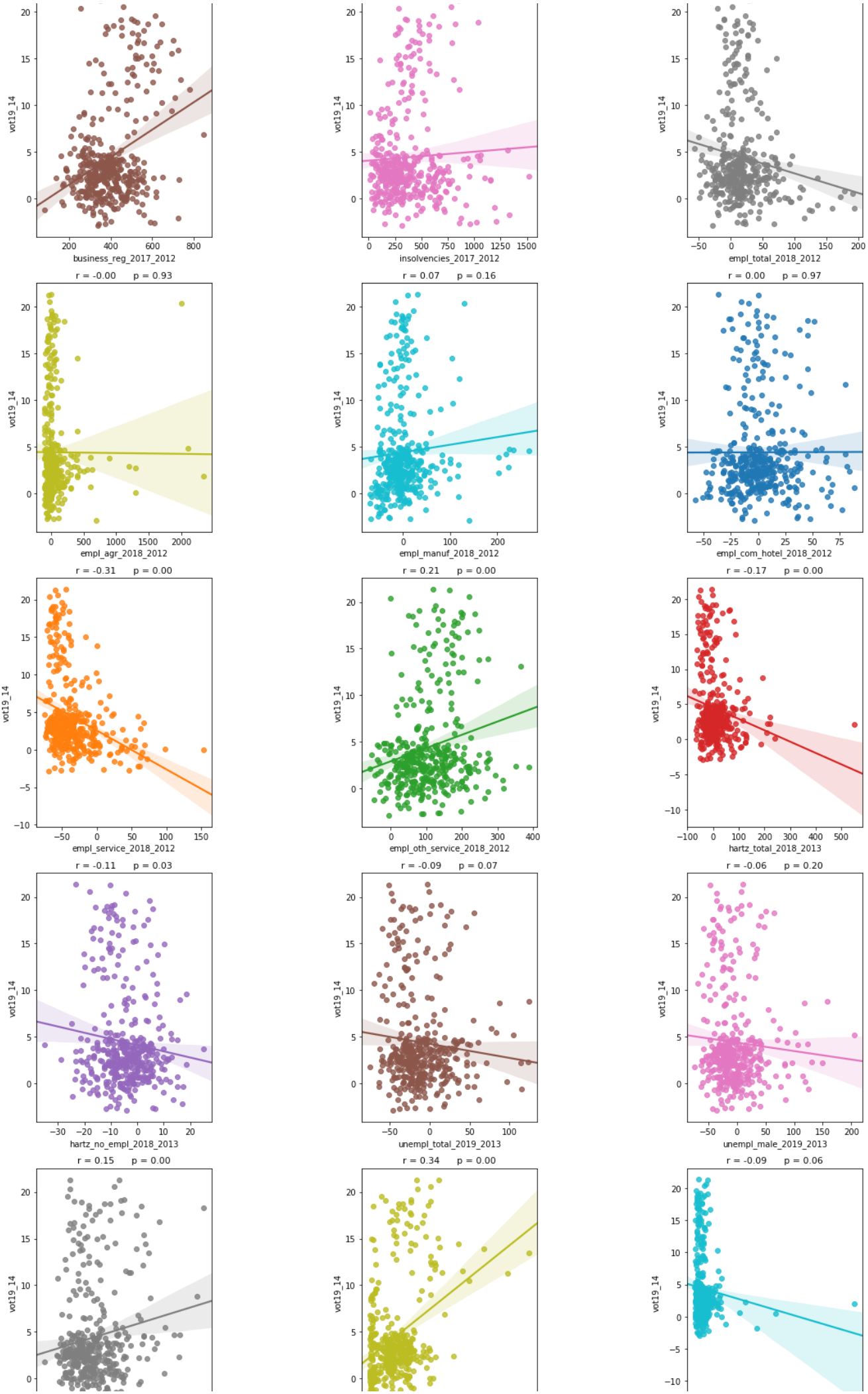

r = 0.17 p = 0.00
vot19_14
dwellings_2017_2012

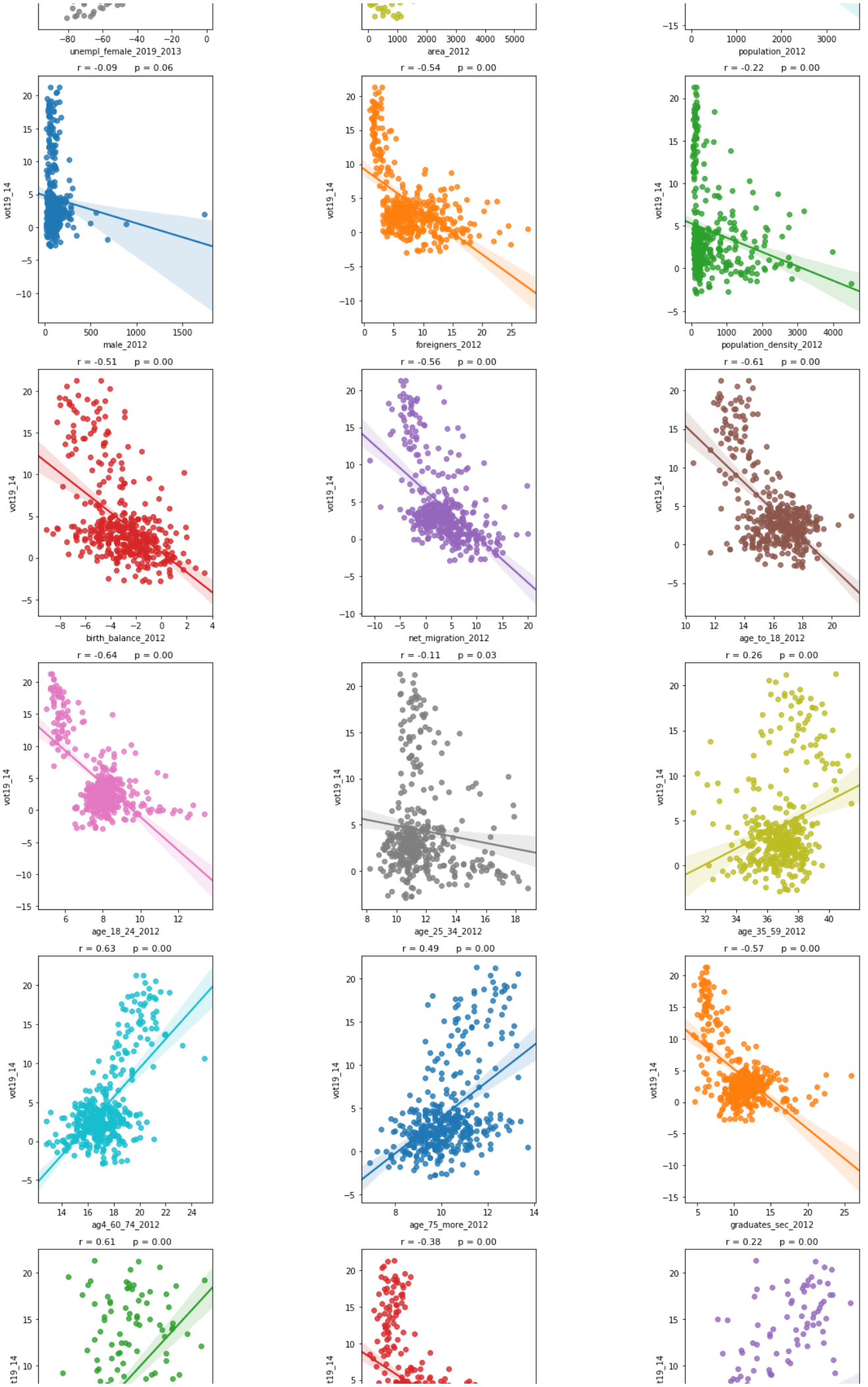

r = 0.07 p = 0.16
vot19_14
vehicles_2018_2013

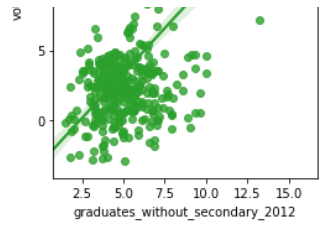

r = 0.33 p = 0.00

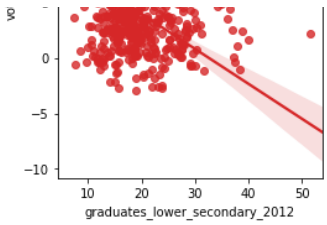

r = 0.05 p = 0.36

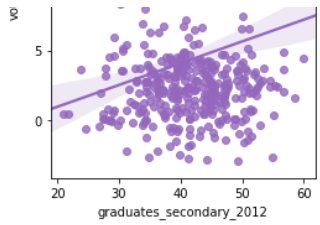

r = -0.14 p = 0.00

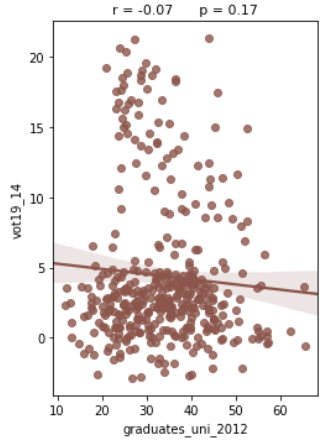

vot19_14
business_reg_2017_2012
insolvencies_2017_2012
empl_total_2018_2012
r = -0.00 p = 0.93
r = 0.07 p = 0.16
r = 0.00 p = 0.97
empl_agr_2018_2012
empl_manuf_2018_2012
empl_com_hotel_2018_2012
r = -0.31 p = 0.00
r = 0.21 p = 0.00
r = -0.17 p = 0.00
empl_service_2018_2012
empl_oth_service_2018_2012
hartz_total_2018_2013
r = -0.11 p = 0.03
r = -0.09 p = 0.07
r = -0.06 p = 0.20
hartz_no_empl_2018_2013
unempl_total_2019_2013
unempl_male_2019_2013
r = 0.15 p = 0.00
r = 0.34 p = 0.00
r = -0.09 p = 0.06

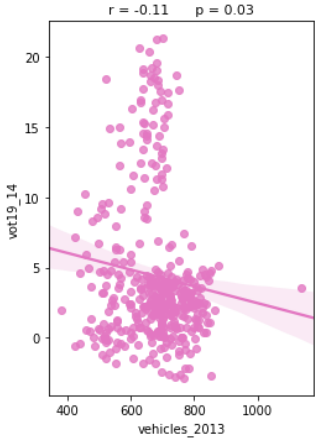

unempl_female_2019_2013
area_2012
population_2012
r = -0.09 p = 0.06
r = -0.54 p = 0.00
r = -0.22 p = 0.00
vot19_14
male_2012
foreigners_2012
population_density_2012
r = -0.51 p = 0.00
r = -0.56 p = 0.00
r = -0.61 p = 0.00
birth_balance_2012
net_migration_2012
age_to_18_2012
r = -0.64 p = 0.00
r = -0.11 p = 0.03
r = 0.26 p = 0.00
age_18_24_2012
age_25_34_2012
age_35_59_2012
r = 0.63 p = 0.00
r = 0.49 p = 0.00
r = -0.57 p = 0.00
ag4_60_74_2012
age_75_more_2012
graduates_sec_2012
r = 0.61 p = 0.00
r = -0.38 p = 0.00
r = 0.22 p = 0.00

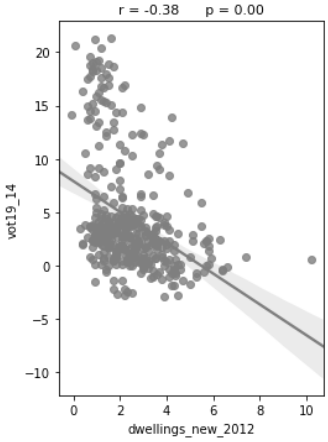

graduates_without_secondary_2012

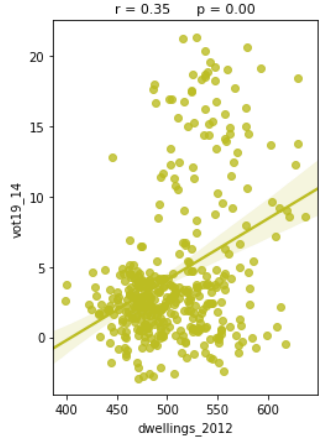

graduates_lower_secondary_2012

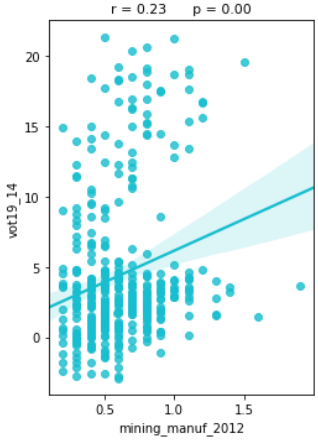

graduates_secondary_2012

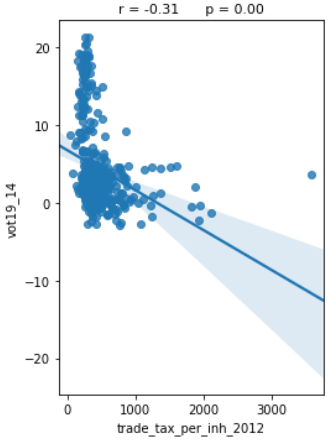

r = -0.07 p = 0.17
vot19_14
graduates_uni_2012

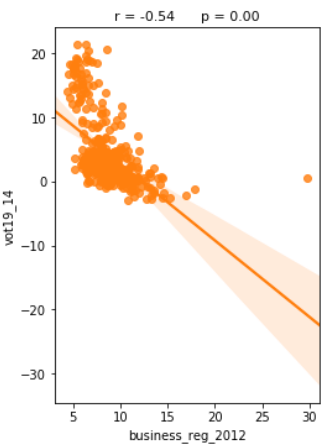

r = -0.11 p = 0.03
vot19_14
vehicles_2013

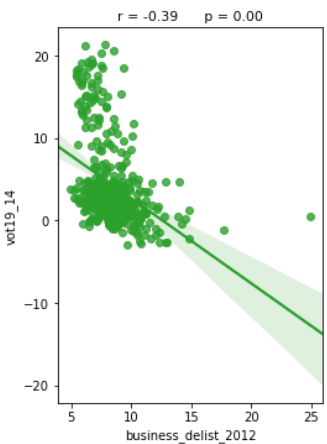

r = -0.38 p = 0.00
vot19_14
dwellings_new_2012

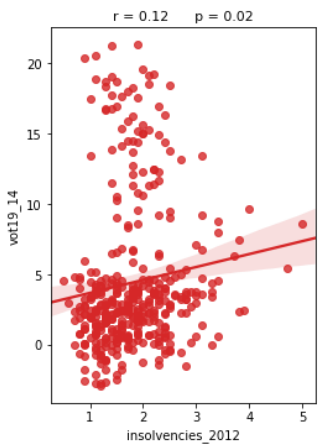

r = 0.35 p = 0.00
vot19_14
dwellings_2012

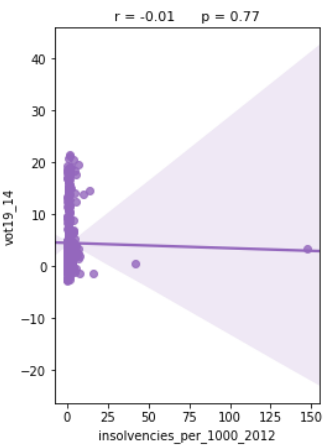

r = 0.23 p = 0.00
vot19_14
mining_manuf_2012

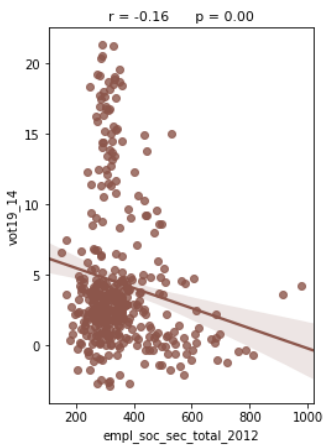

r = -0.31 p = 0.00
vot19_14
trade_tax_per_inh_2012

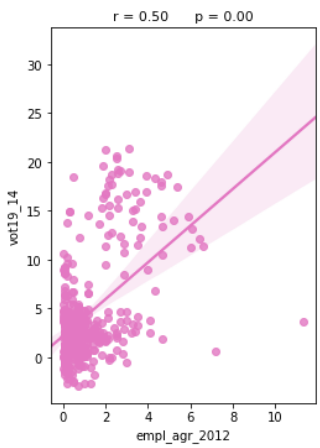

r = -0.54 p = 0.00
vot19_14
business_reg_2012

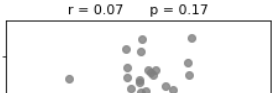

r = -0.39 p = 0.00
vot19_14
business_delist_2012

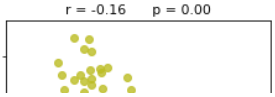

r = 0.12 p = 0.02
vot19_14
insolvencies_2012

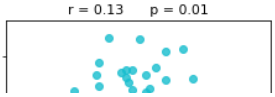

r = -0.01 p = 0.77
vot19_14
insolvencies_per_1000_2012

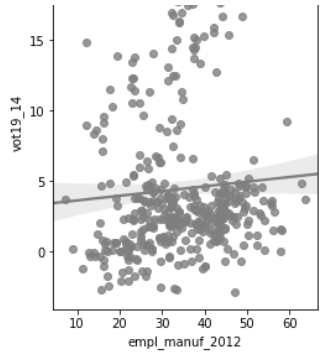

r = -0.16 p = 0.00
vot19_14
empl_soc_sec_total_2012

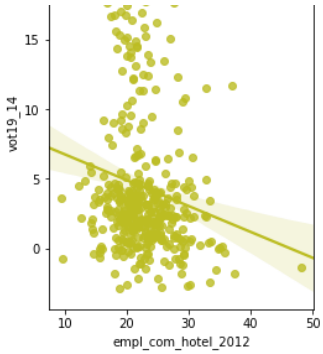

r = 0.50 p = 0.00
vot19_14
empl_agr_2012

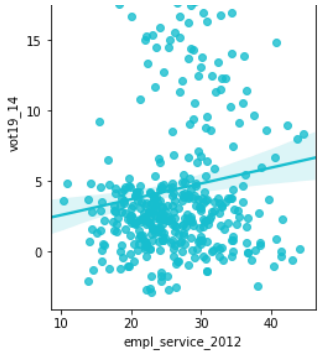

r = 0.07 p = 0.17

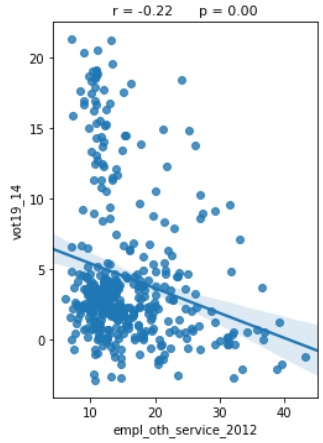

r = -0.16 p = 0.00

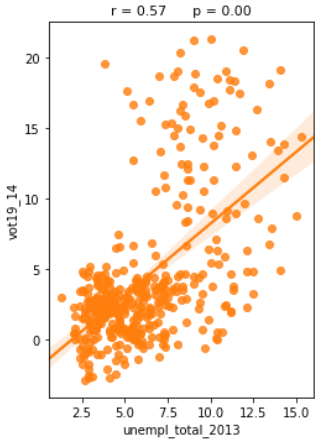

r = 0.13 p = 0.01

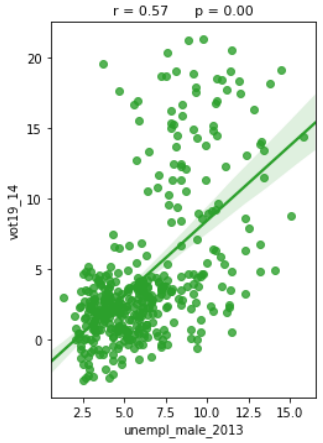

vot19_14
empl_manuf_2012

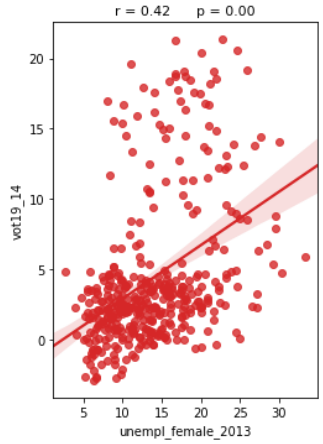

vot19_14
empl_com_hotel_2012

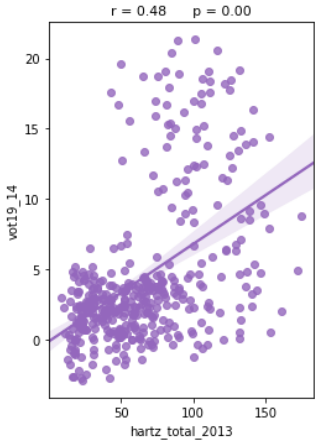

vot19_14
empl_service_2012

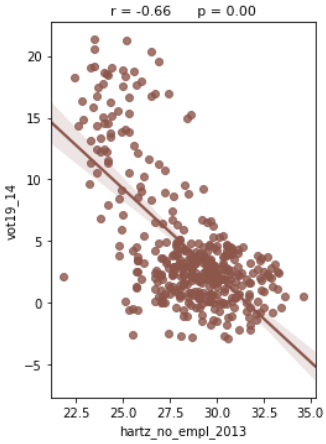

r = -0.22 p = 0.00
vot19_14
empl_oth_service_2012

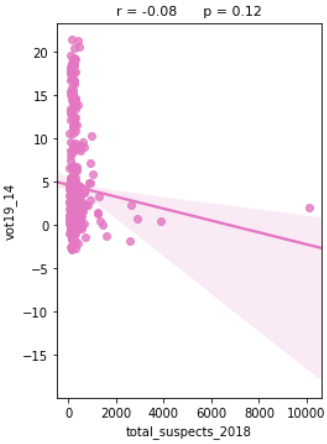

r = 0.57 p = 0.00
vot19_14
unempl_total_2013

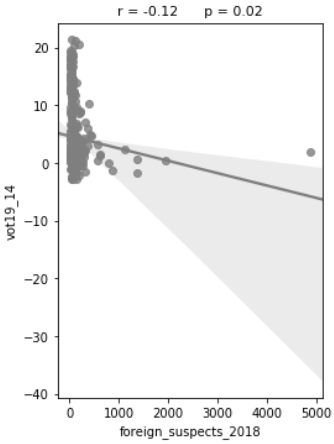

r = 0.57 p = 0.00
vot19_14
unempl_male_2013

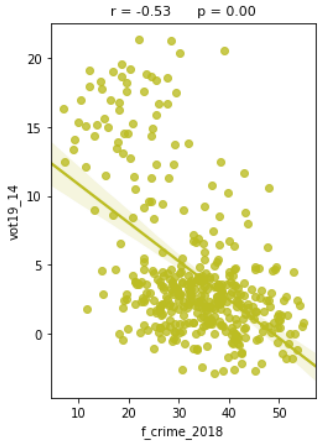

r = 0.42 p = 0.00
vot19_14
unempl_female_2013

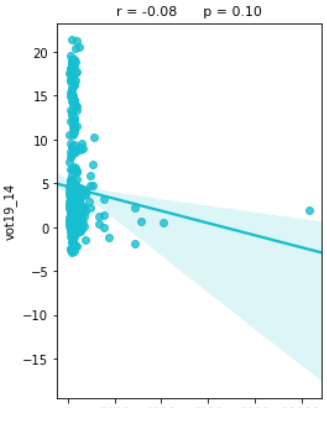

r = 0.48 p = 0.00
vot19_14
hartz_total_2013

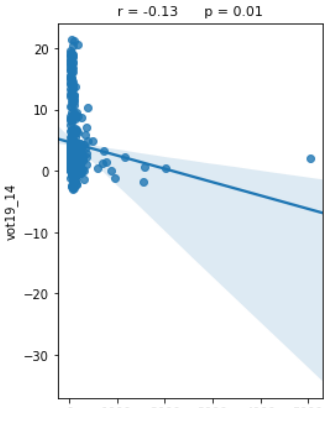

r = -0.66 p = 0.00
vot19_14
hartz_no_empl_2013

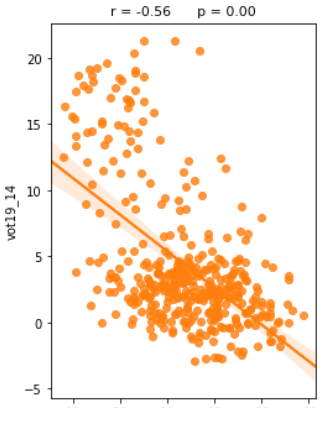

r = -0.08 p = 0.12
vot19_14
total_suspects_2018

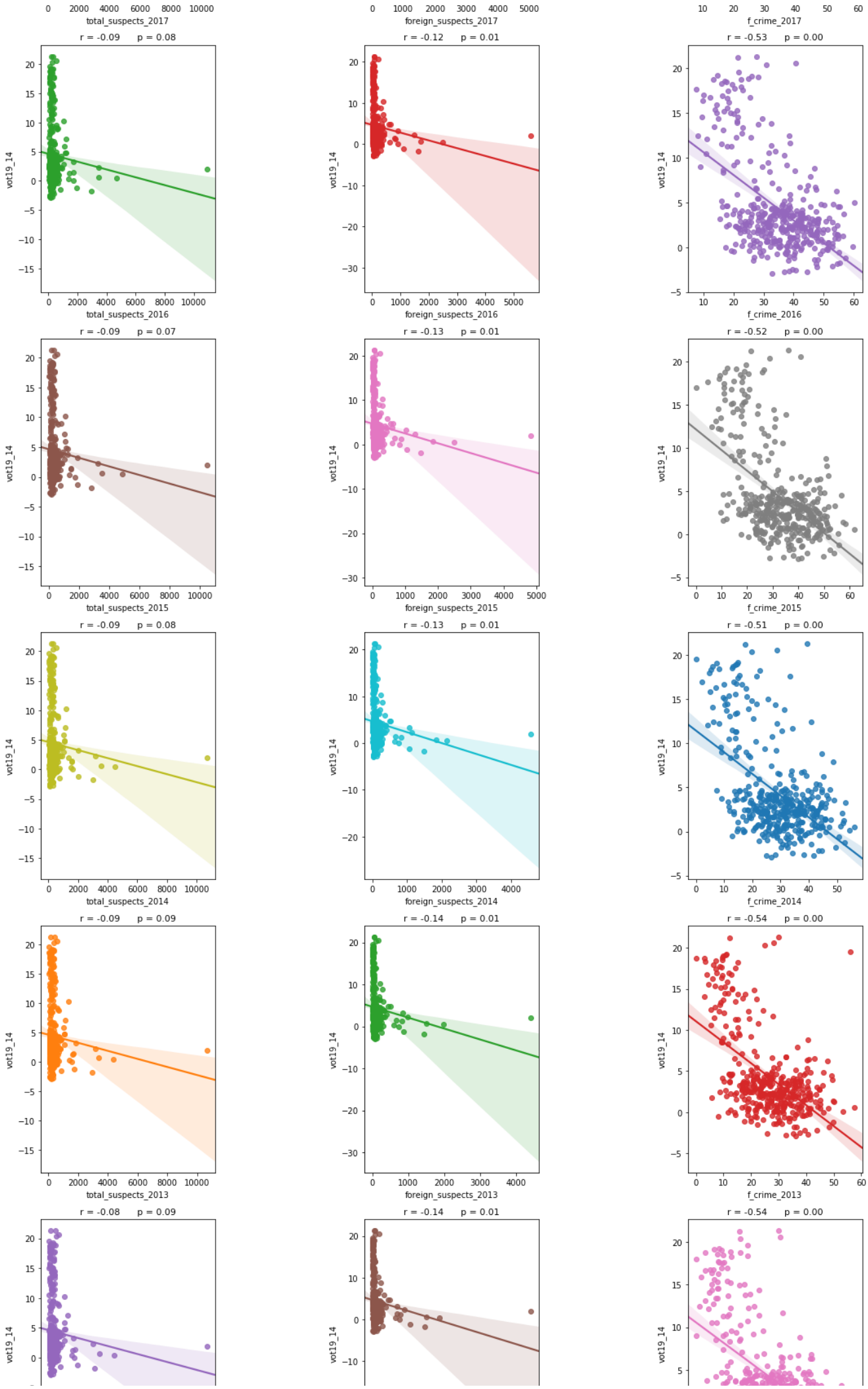

r = -0.12 p = 0.02
vot19_14
foreign_suspects_2018

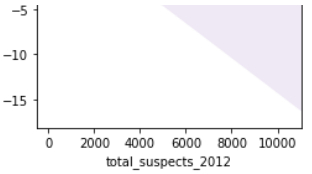

r = -0.53 p = 0.00
vot19_14
f_crime_2018

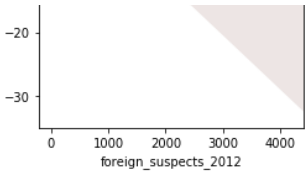

r = -0.08 p = 0.10
vot19_14

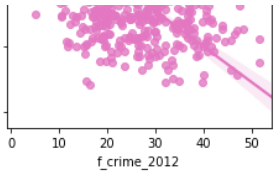

r = -0.13 p = 0.01
vot19_14

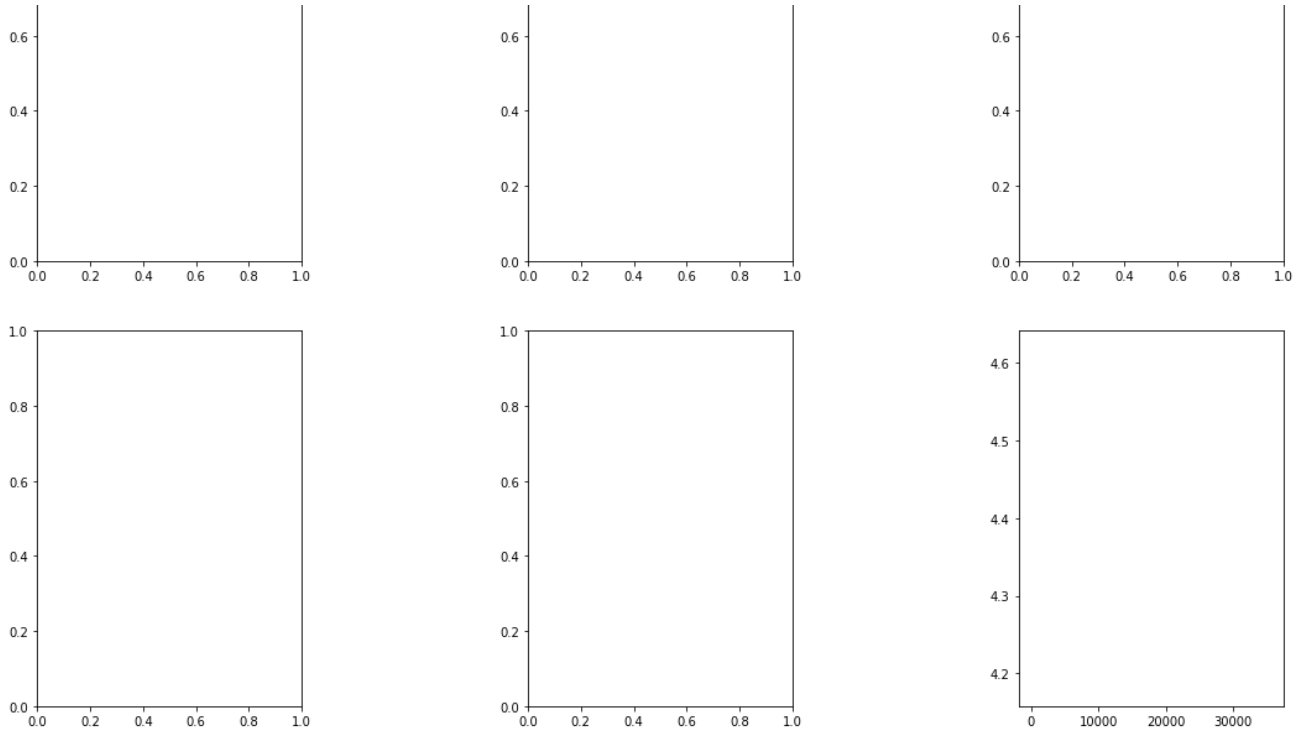

r = -0.56 p = 0.00
vot19_14


total_suspects_2017
r = -0.09 p = 0.08
vot19_14
total_suspects_2016
foreign_suspects_2017
r = -0.12 p = 0.01
foreign_suspects_2016
f_crime_2017
r = -0.53 p = 0.00
f_crime_2016
total_suspects_2016
r = -0.09 p = 0.07
total_suspects_2015
foreign_suspects_2016
r = -0.13 p = 0.01
foreign_suspects_2015
f_crime_2016
r = -0.52 p = 0.00
f_crime_2015
total_suspects_2015
r = -0.09 p = 0.08
total_suspects_2014
foreign_suspects_2015
r = -0.13 p = 0.01
foreign_suspects_2014
f_crime_2015
r = -0.51 p = 0.00
f_crime_2014
total_suspects_2014
r = -0.09 p = 0.09
total_suspects_2013
foreign_suspects_2014
r = -0.14 p = 0.01
foreign_suspects_2013
f_crime_2014
r = -0.54 p = 0.00
f_crime_2013
total_suspects_2013
r = -0.08 p = 0.09
foreign_suspects_2013
r = -0.14 p = 0.01
f_crime_2013
r = -0.54 p = 0.00


total_suspects_2012


foreign_suspects_2012


f_crime_2012



- We see that for some features like 'age4_60_74_2017' there is a strong linear correlation (0.72) to the target.
- the features like 'age35_59_2017' the correlation is very weak.
- For this kernel I decided to use only those features for prediction that have a correlation larger than a threshold value to dependent variable.
- This threshold value can be choosen in the coming steps.

In [16]:

```
target = 'vot19_14'
for col in numerical_feats:
    print('{:15}'.format(col),
          'Skewness: {:05.2f}'.format(df1[col].skew()) ,
          '   ' ,
          'Kurtosis: {:06.2f}'.format(df1[col].kurt())
         )
```

```
Nr              Skewness: 00.34     Kurtosis: -00.28
subregion       Skewness: 00.42     Kurtosis: -00.28
vot19_14        Skewness: 01.51     Kurtosis: 001.47
turnout14       Skewness: -0.02     Kurtosis: -00.28
turnout19       Skewness: -0.09     Kurtosis: -00.16
turnout19_14    Skewness: 00.18     Kurtosis: -00.96
debt_2013       Skewness: 00.79     Kurtosis: 000.94
debt_2014       Skewness: 00.80     Kurtosis: 000.95
debt_2015       Skewness: 00.85     Kurtosis: 001.09
debt_2016       Skewness: 00.84     Kurtosis: 001.02
debt_2017       Skewness: 00.89     Kurtosis: 001.10
debt_2018       Skewness: 00.91     Kurtosis: 001.20
ove18_13        Skewness: 00.27     Kurtosis: 000.76
area_2017       Skewness: 01.71     Kurtosis: 006.23
population_2017 Skewness: 08.51     Kurtosis: 104.70
germans_2017    Skewness: 08.34     Kurtosis: 102.57
foreigners_2017 Skewness: 00.90     Kurtosis: 001.32
population_density_2017 Skewness: 02.31     Kurtosis: 006.15
birth_balance_2017 Skewness: -0.21     Kurtosis: -00.14
net_migration_2017 Skewness: 00.20     Kurtosis: 002.20
age_to_18_2017  Skewness: -0.27     Kurtosis: 000.07
age_18_24_2017  Skewness: 00.18     Kurtosis: 001.58
age_25_34_2017  Skewness: 01.47     Kurtosis: 002.00
age_35_59_2017  Skewness: -0.90     Kurtosis: 001.10
ag4_60_74_2017  Skewness: 00.38     Kurtosis: 000.20
age_75_more_2017 Skewness: 00.56     Kurtosis: 000.27
disposable_inc_2016 Skewness: 00.97     Kurtosis: 002.91
gdp_2016        Skewness: 03.48     Kurtosis: 020.46
protection_total_2017 Skewness: 04.01     Kurtosis: 039.29
protection_open_2017 Skewness: 00.75     Kurtosis: 001.01
```

protection_open_2017 Skewness: 00.75 Kurtosis: 001.01
protection_accepted_2017 Skewness: -0.63 Kurtosis: 000.95
protection_rejected_2017 Skewness: 01.29 Kurtosis: 002.11
dwellings_new_2017 Skewness: 00.78 Kurtosis: 000.17
dwellings_2017 Skewness: 00.52 Kurtosis: -00.23
space_per_app_2017 Skewness: -0.16 Kurtosis: -01.02
space_per_inh_2017 Skewness: 00.33 Kurtosis: 000.62
vehicles_2018 Skewness: -0.29 Kurtosis: 000.68
graduates_voc_2017 Skewness: 01.78 Kurtosis: 004.18
Absolventen/Abgänger allgemeinbildender Schulen 2017 - insgesamt ohne Externe (je 1000 Einwohner)
Skewness: 01.58 Kurtosis: 004.79
graduates_without_secondary_2017 Skewness: 00.80 Kurtosis: 001.00
graduates_lower_secondary_2017 Skewness: 00.60 Kurtosis: 000.61
graduates_secondary_2017 Skewness: -0.23 Kurtosis: -00.35
graduates_higher_2017 Skewness: 00.33 Kurtosis: 000.12
child_day_care_2018 Skewness: 00.93 Kurtosis: -00.13
business_reg_2017 Skewness: 01.15 Kurtosis: 004.12
insolvencies_2017 Skewness: 00.37 Kurtosis: 000.10
empl_total_2018 Skewness: 01.76 Kurtosis: 004.30
empl_agr_2018 Skewness: 02.14 Kurtosis: 006.60
empl_manuf_2018 Skewness: 00.10 Kurtosis: -00.51
empl_com_hotel_2018 Skewness: 00.88 Kurtosis: 002.46
empl_service_2018 Skewness: 01.44 Kurtosis: 002.07
empl_oth_service_2018 Skewness: 00.18 Kurtosis: -00.37
hartz_total_2018 Skewness: 00.95 Kurtosis: 000.68
hartz_no_empl_2018 Skewness: -0.43 Kurtosis: -00.24
hartz_foreign_2018 Skewness: -0.65 Kurtosis: -00.31
unempl_total_2019 Skewness: 00.91 Kurtosis: 000.53
unempl_male_2019 Skewness: 00.84 Kurtosis: 000.36
unempl_female_2019 Skewness: 00.96 Kurtosis: 000.78
unempl_15_19_2019 Skewness: 01.13 Kurtosis: 000.87
unempl_55_64_2019 Skewness: 00.87 Kurtosis: 000.62
foreigners_2017_2012 Skewness: 02.82 Kurtosis: 011.26
population_density_2017_2012 Skewness: 05.62 Kurtosis: 046.70
birth_balance_2017_2012 Skewness: -1.14 Kurtosis: 059.14
net_migration_2017_2012 Skewness: 01.74 Kurtosis: 062.35
age_to_18_2017_2012 Skewness: 00.89 Kurtosis: 001.55
age_18_24_2017_2012 Skewness: 00.91 Kurtosis: 002.82
age_25_34_2017_2012 Skewness: 01.06 Kurtosis: 002.79
age_35_59_2017_2012 Skewness: -0.72 Kurtosis: 001.45
ag4_60_74_2017_2012 Skewness: 00.20 Kurtosis: 001.21
age_75_more_2017_2012 Skewness: 00.48 Kurtosis: 000.52
graduates_without_secondary_2017_2012 Skewness: 04.62 Kurtosis: 041.48
graduates_lower_secondary_2017_2012 Skewness: 01.24 Kurtosis: 003.54
graduates_secondary_2017_2012 Skewness: 01.52 Kurtosis: 005.80
dwellings_new_2017_2012 Skewness: 01.94 Kurtosis: 038.91
dwellings_2017_2012 Skewness: 00.48 Kurtosis: 000.90
vehicles_2018_2013 Skewness: 00.91 Kurtosis: 004.13
business_reg_2017_2012 Skewness: 00.57 Kurtosis: 000.55
insolvencies_2017_2012 Skewness: 01.24 Kurtosis: 001.88
empl_total_2018_2012 Skewness: 01.59 Kurtosis: 004.69
empl_agr_2018_2012 Skewness: 05.86 Kurtosis: 042.70
empl_manuf_2018_2012 Skewness: 02.46 Kurtosis: 009.42
empl_com_hotel_2018_2012 Skewness: 00.78 Kurtosis: 001.05
empl_service_2018_2012 Skewness: 01.88 Kurtosis: 004.87
empl_oth_service_2018_2012 Skewness: 00.49 Kurtosis: 000.31
hartz_total_2018_2013 Skewness: 03.32 Kurtosis: 022.95
hartz_no_empl_2018_2013 Skewness: 00.05 Kurtosis: 000.31
unempl_total_2019_2013 Skewness: 01.26 Kurtosis: 002.59
unempl_male_2019_2013 Skewness: 01.65 Kurtosis: 004.55
unempl_female_2019_2013 Skewness: 01.30 Kurtosis: 002.48
area_2012 Skewness: 01.70 Kurtosis: 006.17
population_2012 Skewness: 08.50 Kurtosis: 104.69
male_2012 Skewness: 08.53 Kurtosis: 105.43
foreigners_2012 Skewness: 01.03 Kurtosis: 001.35
population_density_2012 Skewness: 02.30 Kurtosis: 006.01
birth_balance_2012 Skewness: -0.02 Kurtosis: -00.31
net_migration_2012 Skewness: 00.34 Kurtosis: 000.34
age_to_18_2012 Skewness: -0.42 Kurtosis: -00.33
age_18_24_2012 Skewness: 00.42 Kurtosis: 001.94
age_25_34_2012 Skewness: 01.55 Kurtosis: 002.25
age_35_59_2012 Skewness: -0.45 Kurtosis: 001.25
ag4_60_74_2012 Skewness: 00.48 Kurtosis: 000.75
age_75_more_2012 Skewness: 00.30 Kurtosis: -00.11
graduates_sec_2012 Skewness: 00.62 Kurtosis: 001.74
graduates_without_secondary_2012 Skewness: 01.29 Kurtosis: 001.84
graduates_lower_secondary_2012 Skewness: 01.08 Kurtosis: 001.90
graduates_secondary_2012 Skewness: -0.13 Kurtosis: -00.43

```
graduates_secondary_2012 Skewness: 0.15     Kurtosis: -00.45
graduates_uni_2012 Skewness: 00.35     Kurtosis: -00.08
vehicles_2013   Skewness: -0.22     Kurtosis: 000.78
dwellings_new_2012 Skewness: 01.16     Kurtosis: 002.30
dwellings_2012  Skewness: 00.47     Kurtosis: -00.05
mining_manuf_2012 Skewness: 00.95     Kurtosis: 001.36
trade_tax_per_inh_2012 Skewness: 04.10     Kurtosis: 027.50
business_reg_2012 Skewness: 02.31     Kurtosis: 015.33
business_delist_2012 Skewness: 02.18     Kurtosis: 012.15
insolvencies_2012 Skewness: 01.00     Kurtosis: 001.73
insolvencies_per_1000_2012 Skewness: 17.05     Kurtosis: 315.28
empl_soc_sec_total_2012 Skewness: 01.86     Kurtosis: 004.85
empl_agr_2012   Skewness: 02.43     Kurtosis: 008.97
empl_manuf_2012 Skewness: 00.06     Kurtosis: -00.49
empl_com_hotel_2012 Skewness: 00.80     Kurtosis: 002.28
empl_service_2012 Skewness: 00.34     Kurtosis: -00.04
empl_oth_service_2012 Skewness: 01.44     Kurtosis: 001.97
unempl_total_2013 Skewness: 00.78     Kurtosis: -00.05
unempl_male_2013 Skewness: 00.81     Kurtosis: 000.09
unempl_female_2013 Skewness: 00.62     Kurtosis: -00.11
hartz_total_2013 Skewness: 00.62     Kurtosis: -00.38
hartz_no_empl_2013 Skewness: -0.35     Kurtosis: -00.36
total_suspects_2018 Skewness: 11.81     Kurtosis: 177.78
foreign_suspects_2018 Skewness: 11.93     Kurtosis: 178.61
f_crime_2018    Skewness: -0.14     Kurtosis: -00.39
total_suspects_2017 Skewness: 11.38     Kurtosis: 166.68
foreign_suspects_2017 Skewness: 11.48     Kurtosis: 167.13
f_crime_2017    Skewness: -0.17     Kurtosis: -00.52
total_suspects_2016 Skewness: 10.81     Kurtosis: 151.68
foreign_suspects_2016 Skewness: 11.01     Kurtosis: 155.13
f_crime_2016    Skewness: -0.12     Kurtosis: -00.61
total_suspects_2015 Skewness: 10.48     Kurtosis: 143.24
foreign_suspects_2015 Skewness: 09.92     Kurtosis: 127.19
f_crime_2015    Skewness: -0.09     Kurtosis: -00.58
total_suspects_2014 Skewness: 10.64     Kurtosis: 148.31
foreign_suspects_2014 Skewness: 10.04     Kurtosis: 130.21
f_crime_2014    Skewness: -0.08     Kurtosis: -00.50
total_suspects_2013 Skewness: 10.97     Kurtosis: 156.40
foreign_suspects_2013 Skewness: 10.64     Kurtosis: 145.52
f_crime_2013    Skewness: 00.05     Kurtosis: -00.56
total_suspects_2012 Skewness: 10.76     Kurtosis: 151.82
foreign_suspects_2012 Skewness: 10.70     Kurtosis: 147.39
f_crime_2012    Skewness: 00.14     Kurtosis: -00.62
```

In [17]:

```
min_val_corr = 0.5
corr = df1.corr()
corr_abs = corr.abs()

numerical_cols = len(numerical_feats)
high_corr = corr_abs.nlargest(numerical_cols, target)[target]

cols_abv_corr_limit = list(high_corr[high_corr.values > min_val_corr].index)
cols_bel_corr_limit = list(high_corr[high_corr.values <= min_val_corr].index)
```

In [32]:

```
print(high_corr)
print("*"*30)
print("List of numerical features with r above min_val_corr :")
print(cols_abv_corr_limit)
print("*"*30)
print("List of numerical features with r below min_val_corr :")
print(cols_bel_corr_limit)
```

```
vot19_14                          1.000000
hartz_foreign_2018                0.768042
age_18_24_2017                    0.733853
child_day_care_2018               0.730496
ag4_60_74_2017                    0.715892
                                    ...
graduates_secondary_2017_2012     0.027159
insolvencies_per_1000_2012        0.014906
empl_agr_2018_2012                0.004121
```

```
dwellings_new_2017_2012          0.003734
empl_com_hotel_2018_2012         0.001781
Name: vot19_14, Length: 148, dtype: float64
******************************
List of numerical features with r above min_val_corr :
['vot19_14', 'hartz_foreign_2018', 'age_18_24_2017', 'child_day_care_2018', 'ag4_60_74_2017', 'sub
region', 'Nr', 'hartz_no_empl_2013', 'age_18_24_2012', 'age_75_more_2017', 'ag4_60_74_2012',
'birth_balance_2017', 'age_to_18_2012', 'graduates_without_secondary_2012', 'hartz_no_empl_2018',
'disposable_inc_2016', 'unempl_male_2013', 'graduates_sec_2012', 'unempl_total_2013',
'net_migration_2012', 'f_crime_2017', 'unempl_55_64_2019', 'foreigners_2017', 'foreigners_2012',
'f_crime_2012', 'business_reg_2012', 'f_crime_2013', 'f_crime_2016', 'f_crime_2018',
'f_crime_2015', 'protection_rejected_2017', 'birth_balance_2012', 'f_crime_2014',
'dwellings_2017']
******************************
List of numerical features with r below min_val_corr :
['empl_agr_2012', 'age_75_more_2012', 'graduates_without_secondary_2017', 'hartz_total_2013', 'une
mpl_male_2019', 'unempl_total_2019', 'unempl_15_19_2019', 'age_18_24_2017_2012',
'unempl_female_2019', 'age_to_18_2017', 'empl_agr_2018', 'Absolventen/Abgänger allgemeinbildender
Schulen 2017 - insgesamt ohne Externe (je 1000 Einwohner)', 'unempl_female_2013',
'business_delist_2012', 'space_per_app_2017', 'graduates_lower_secondary_2012',
'dwellings_new_2012', 'dwellings_new_2017', 'business_reg_2017', 'insolvencies_2017',
'age_25_34_2017_2012', 'turnout19_14', 'age_to_18_2017_2012', 'dwellings_2012', 'age_25_34_2017',
'area_2017', 'area_2012', 'graduates_voc_2017', 'net_migration_2017', 'business_reg_2017_2012', 'g
dp_2016', 'graduates_lower_secondary_2017', 'empl_service_2018_2012', 'trade_tax_per_inh_2012', 'a
ge_75_more_2017_2012', 'hartz_total_2018', 'protection_accepted_2017', 'age_35_59_2012',
'ag4_60_74_2017_2012', 'empl_service_2018', 'ove18_13', 'mining_manuf_2012',
'population_density_2017', 'empl_oth_service_2012', 'population_density_2012',
'graduates_secondary_2012', 'empl_oth_service_2018_2012', 'debt_2018', 'debt_2017', 'debt_2016', '
debt_2015', 'foreigners_2017_2012', 'empl_total_2018', 'graduates_lower_secondary_2017_2012',
'debt_2014', 'turnout19', 'protection_total_2017', 'hartz_total_2018_2013', 'dwellings_2017_2012',
'empl_com_hotel_2012', 'empl_oth_service_2018', 'debt_2013', 'empl_soc_sec_total_2012',
'space_per_inh_2017', 'unempl_female_2019_2013', 'empl_total_2018_2012', 'foreign_suspects_2012',
'foreign_suspects_2013', 'graduates_secondary_2017', 'empl_service_2012', 'foreign_suspects_2015',
'foreign_suspects_2014', 'age_35_59_2017_2012', 'turnout14', 'foreign_suspects_2017',
'foreign_suspects_2016', 'empl_com_hotel_2018', 'insolvencies_2012', 'foreign_suspects_2018',
'age_25_34_2012', 'hartz_no_empl_2018_2013', 'vehicles_2013', 'vehicles_2018', 'population_2017',
'birth_balance_2017_2012', 'empl_manuf_2018', 'population_2012', 'male_2012',
'unempl_total_2019_2013', 'total_suspects_2015', 'graduates_without_secondary_2017_2012',
'total_suspects_2014', 'total_suspects_2016', 'total_suspects_2013', 'total_suspects_2012',
'total_suspects_2017', 'germans_2017', 'total_suspects_2018', 'empl_manuf_2018_2012',
'vehicles_2018_2013', 'graduates_uni_2012', 'population_density_2017_2012', 'empl_manuf_2012', 'un
empl_male_2019_2013', 'protection_open_2017', 'graduates_higher_2017', 'net_migration_2017_2012',
'insolvencies_2017_2012', 'age_35_59_2017', 'graduates_secondary_2017_2012',
'insolvencies_per_1000_2012', 'empl_agr_2018_2012', 'dwellings_new_2017_2012',
'empl_com_hotel_2018_2012']
```

In [33]:

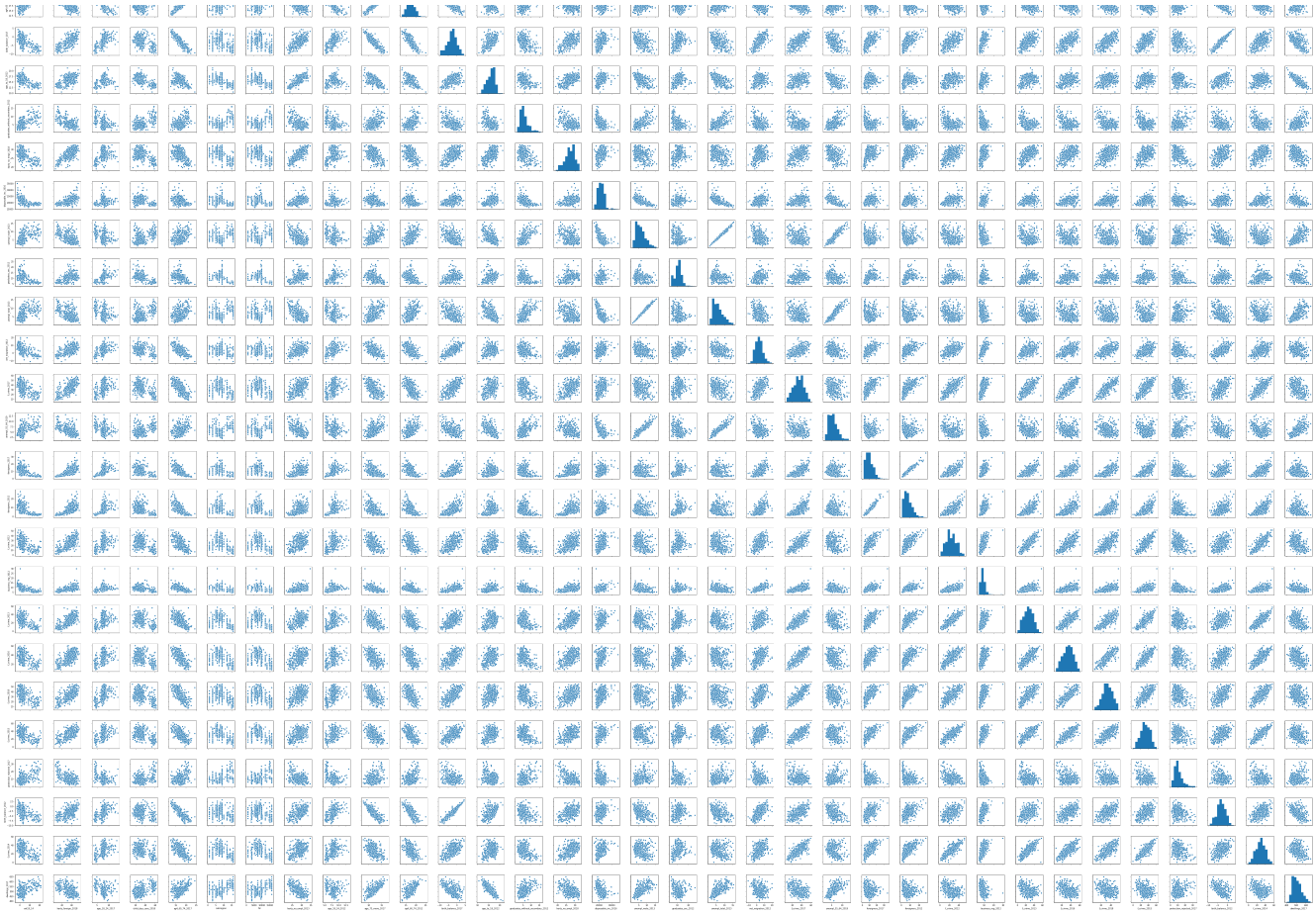
```
df2 = df1[cols_abv_corr_limit]
corr = df2.corr()
```

In [15]:

```
sns.pairplot(df2)
```

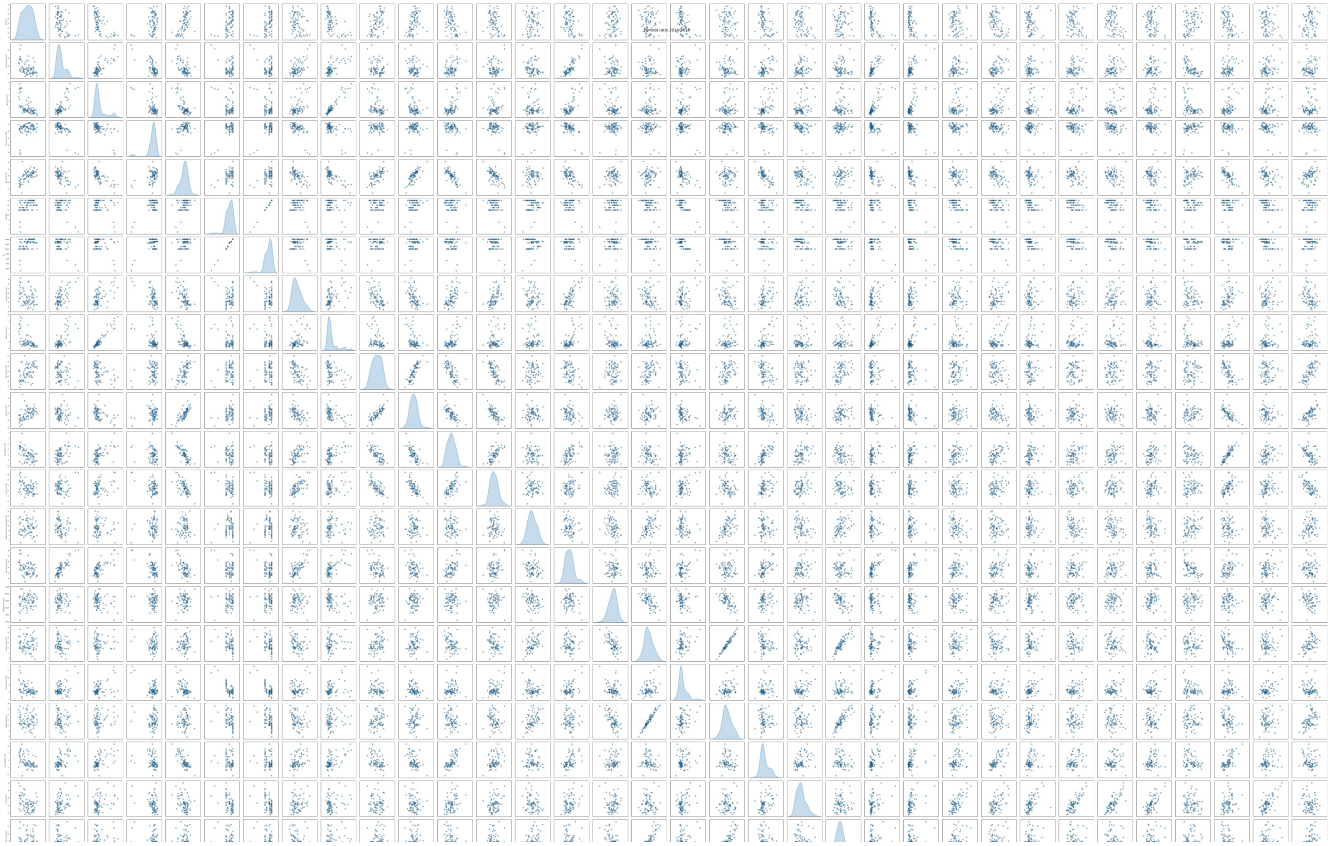
Out[15]:

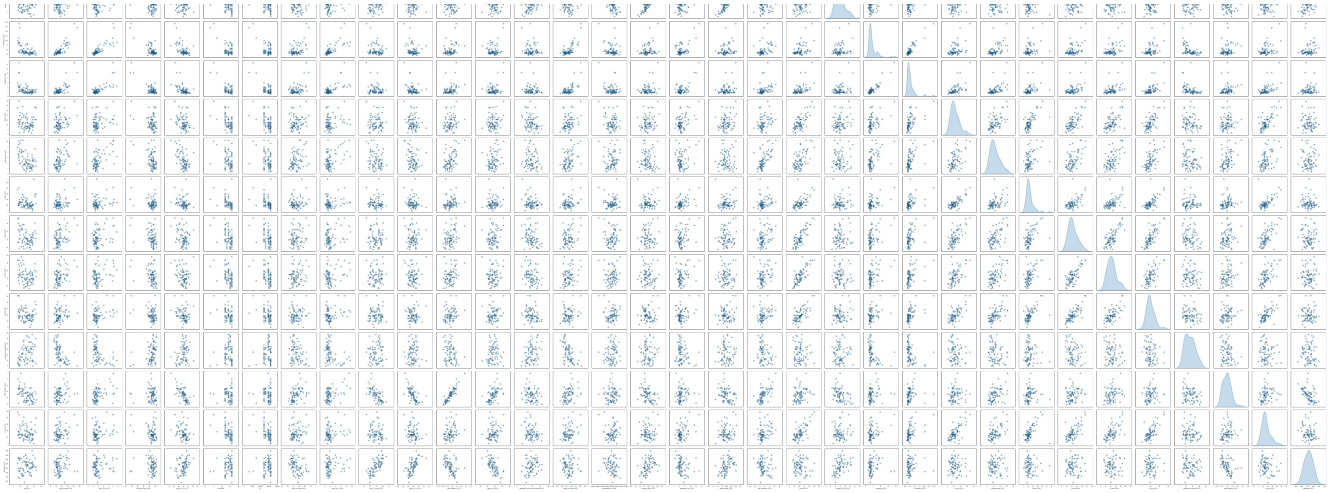
```
<seaborn.axisgrid.PairGrid at 0x298af872708>
```



In [21]:

```
# Plot colored by continent for years 2000-2007
sns.pairplot(df2[df2['vot19_14'] >= 7.9],  diag_kind = 'kde',
             plot_kws = {'alpha': 0.6, 's': 80, 'edgecolor': 'k'},
             size = 4);
# Title
plt.suptitle('Turnout ratio 2014-2019',
             size = 28);
```



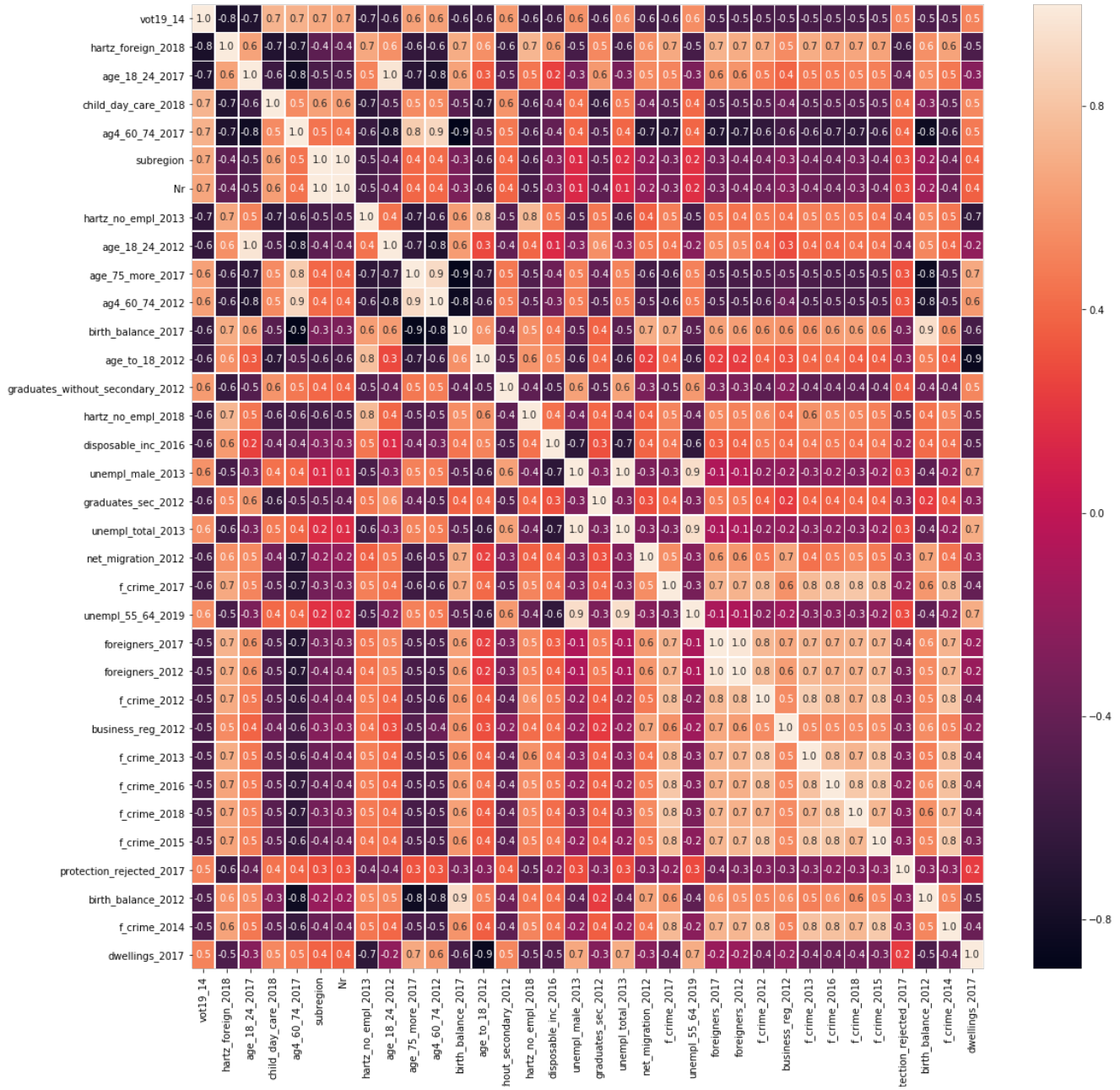


In [178]:

```
f,ax = plt.subplots(figsize=(18, 18))
sns.heatmap(df2.corr(), annot=True, linewidths=.5, fmt= '.1f',ax=ax)
```

Out[178]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1f5f86634a8>
```



graduates_wit

prol

- As the linear regression has the tendency to get overfitted with data having collinearity between independent varaibles, while fitting the data to a model we need to deal with multicolinearity between them .
- So in the forth coming feature selection and extraction steps we will deal with that using some feature selection techniques.
- From the above analysis with respect to target Region,state abbrev have been eliminated by the corr matrix as they dont really constitute to the target variable so be remove them from the data set before building our model

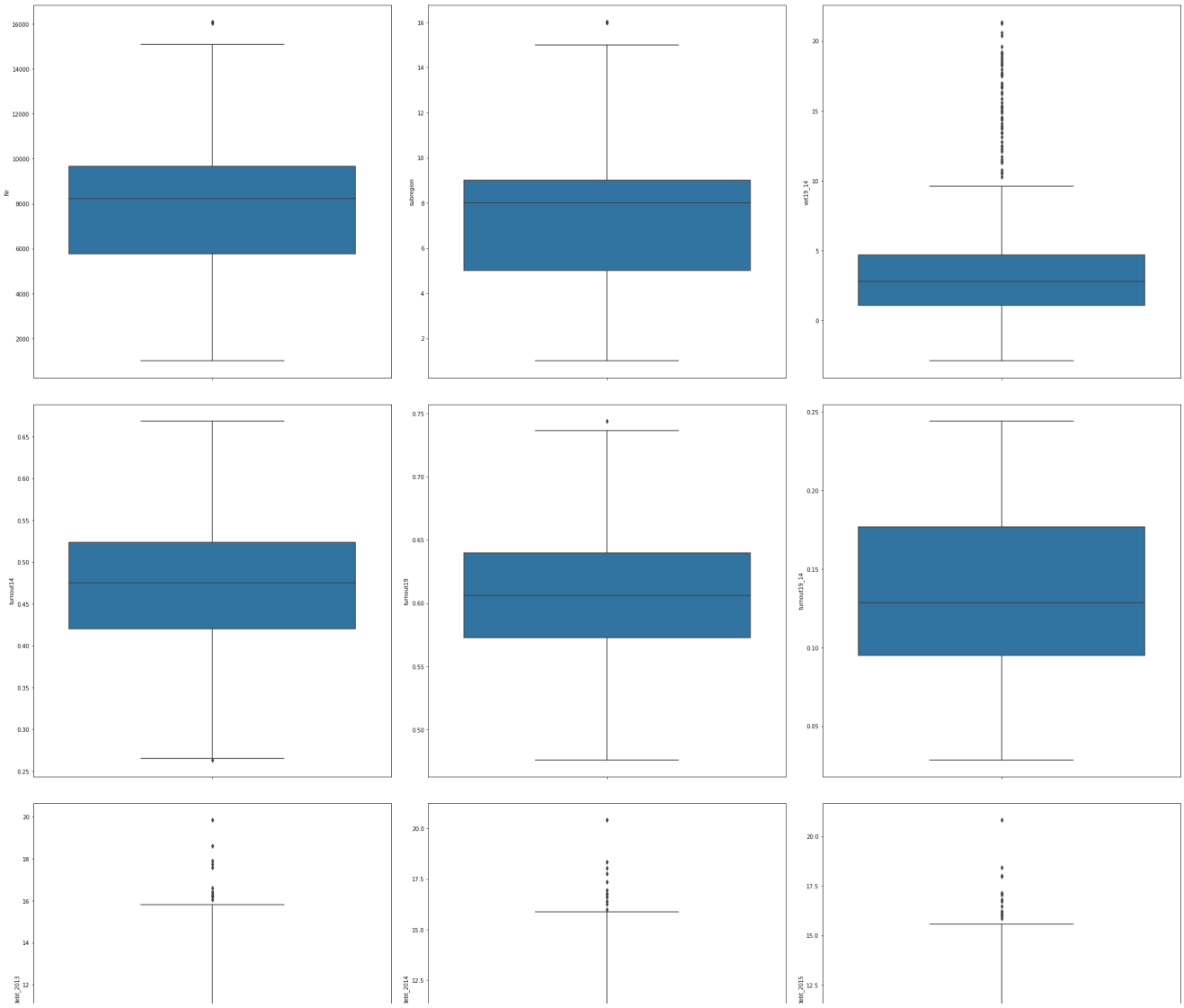
In [50]:

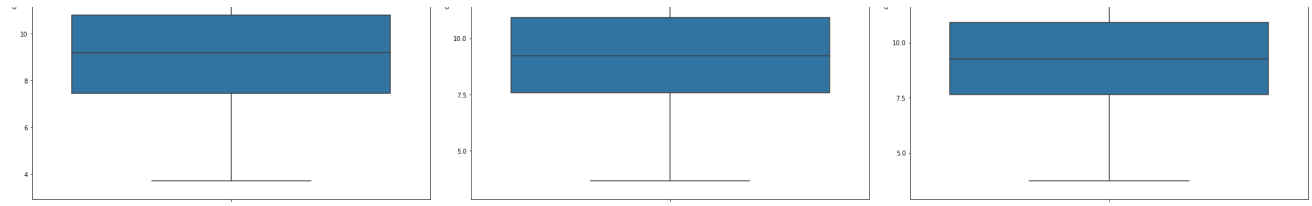
```
df1.drop(columns = {'region','state_abbrev','state'},axis = 1,inplace = True)
```

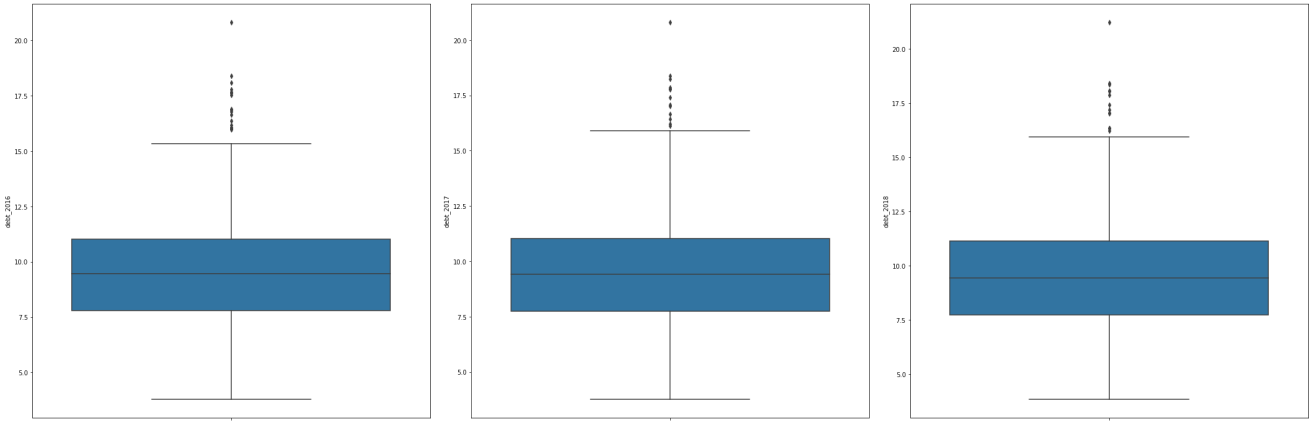
Lets visualize outliers using boxplots and find the no"s in the data

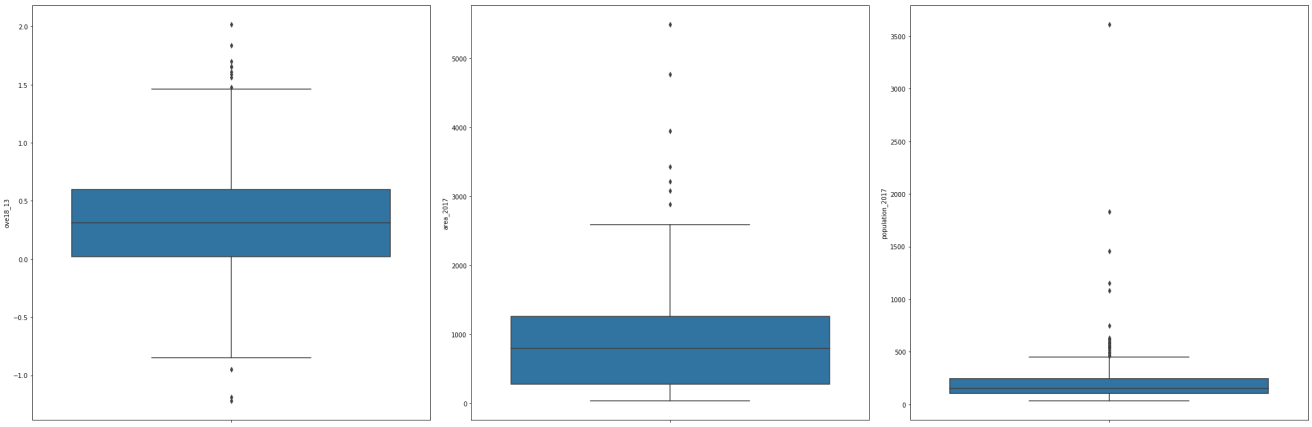
In [185]:

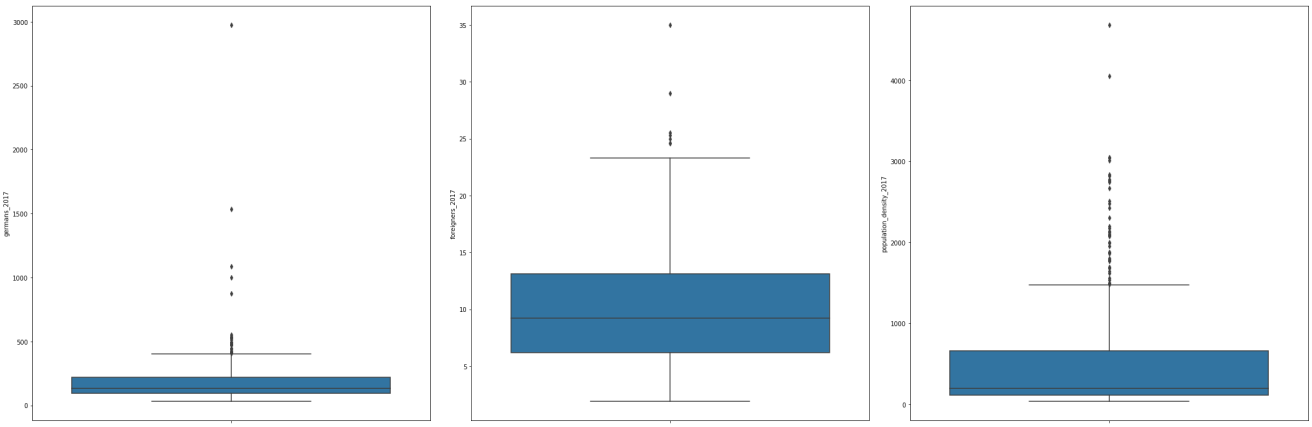
```
# from the above we find diff in min and max values so lets visualize the oultiers uisng Box plot
of  all the Attributes.
n_columns = 3
n_rows = 55
f, axes = plt.subplots(n_rows, n_columns, figsize=(10* n_columns, 10* n_rows))
for i, c in enumerate(df1.columns):
    sns.boxplot(y = c, data = df1, ax = axes[i // n_columns, i % n_columns])
plt.tight_layout()
plt.show()
```

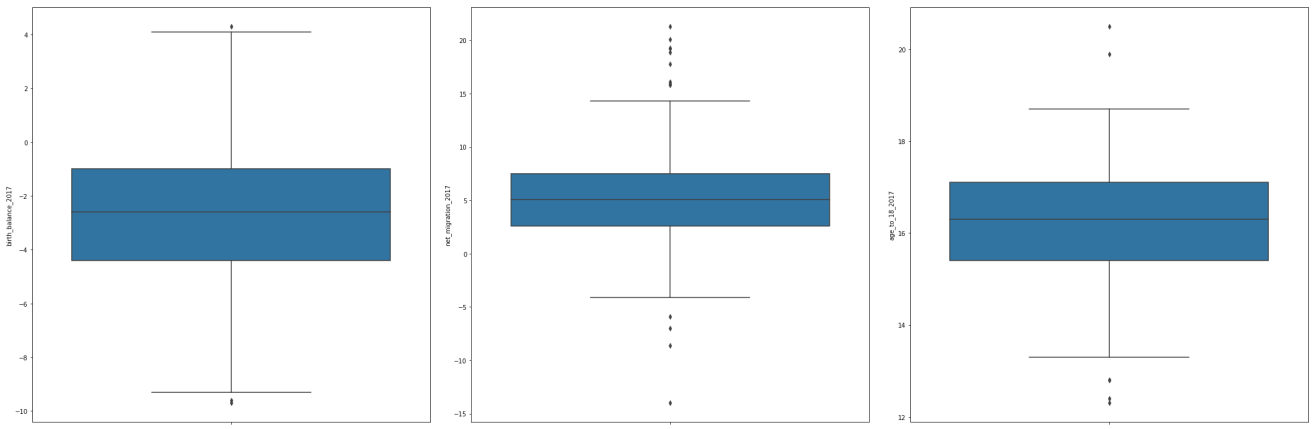


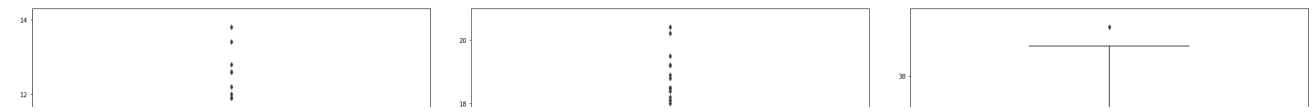


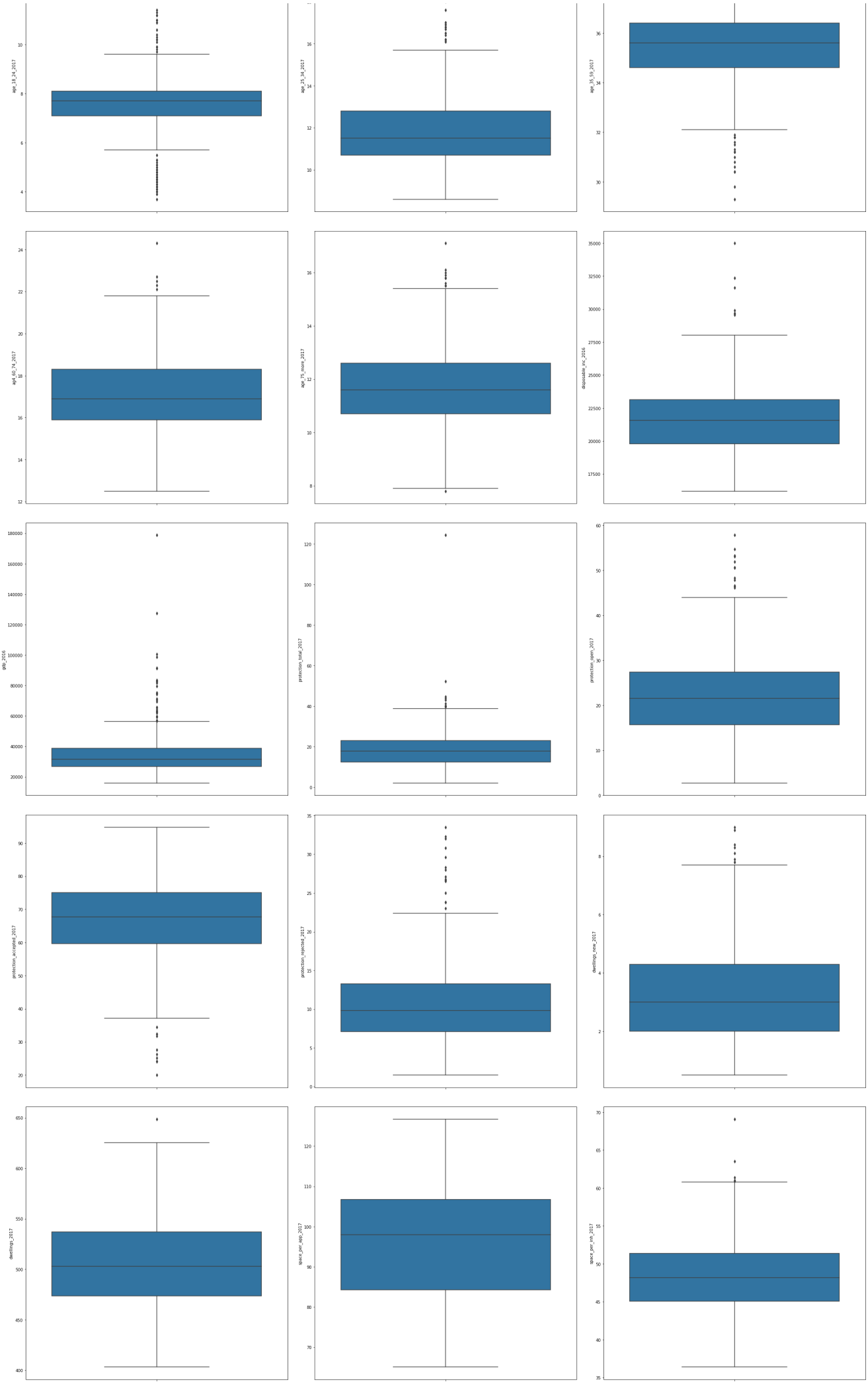

debt_2016
debt_2017
debt_2018

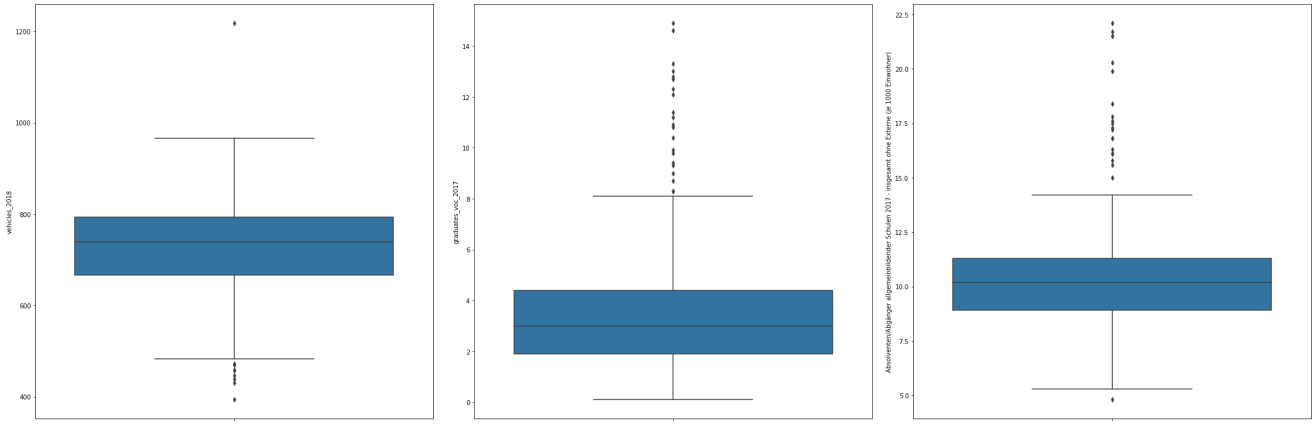

ovc18_13
area_2017
population_2017

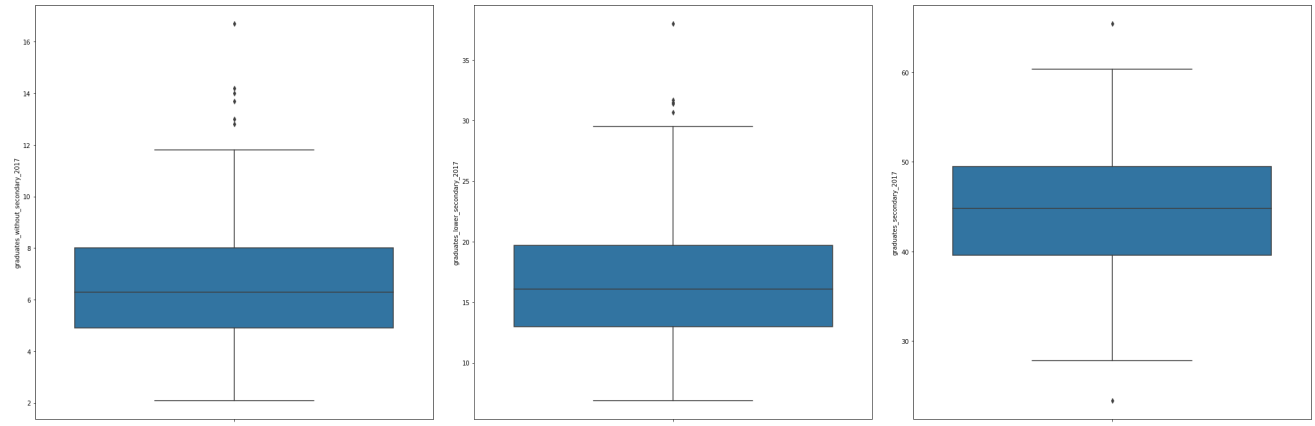

germans_2017
foreigners_2017
population_density_2017

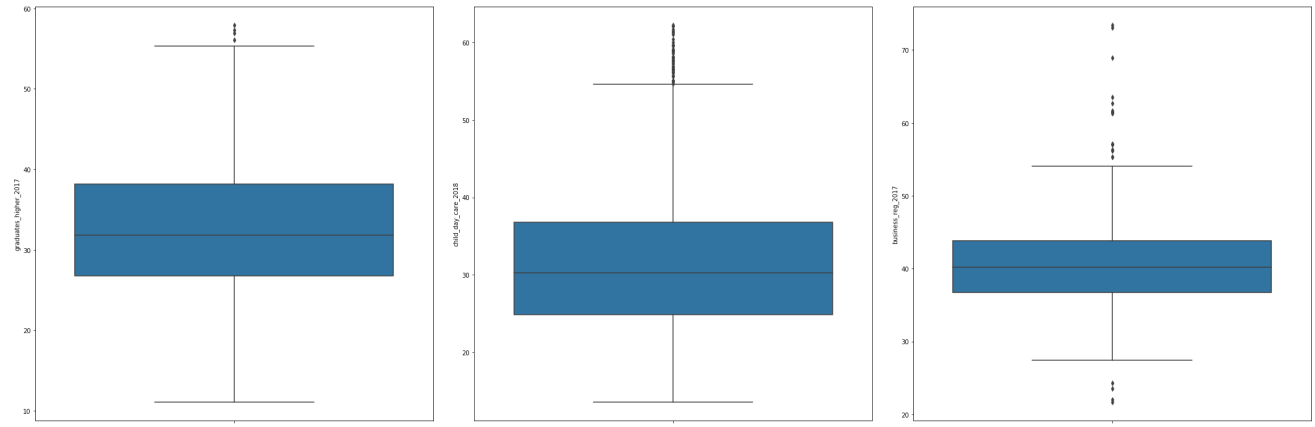

birth_balance_2017
net_migration_2017
age_to_18_2017

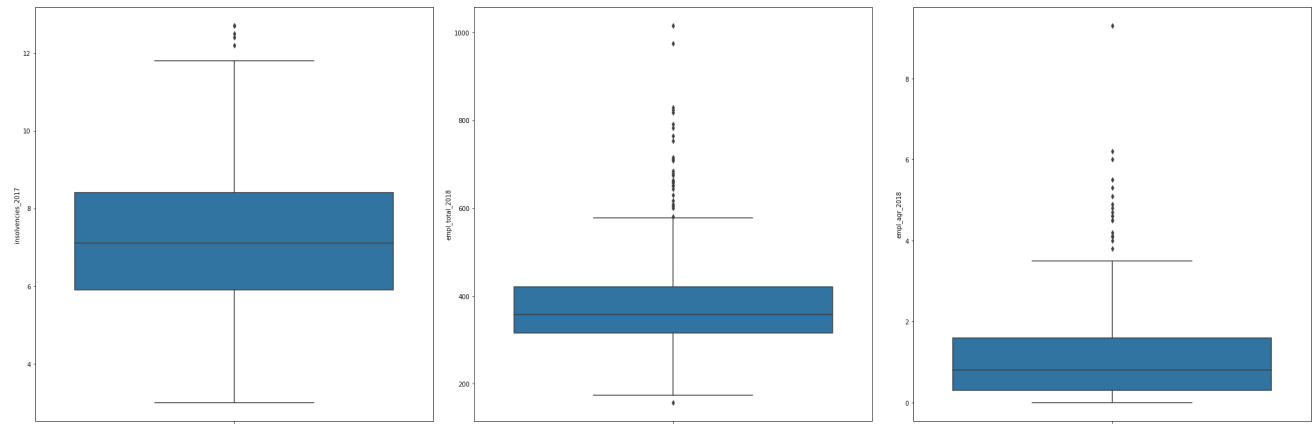


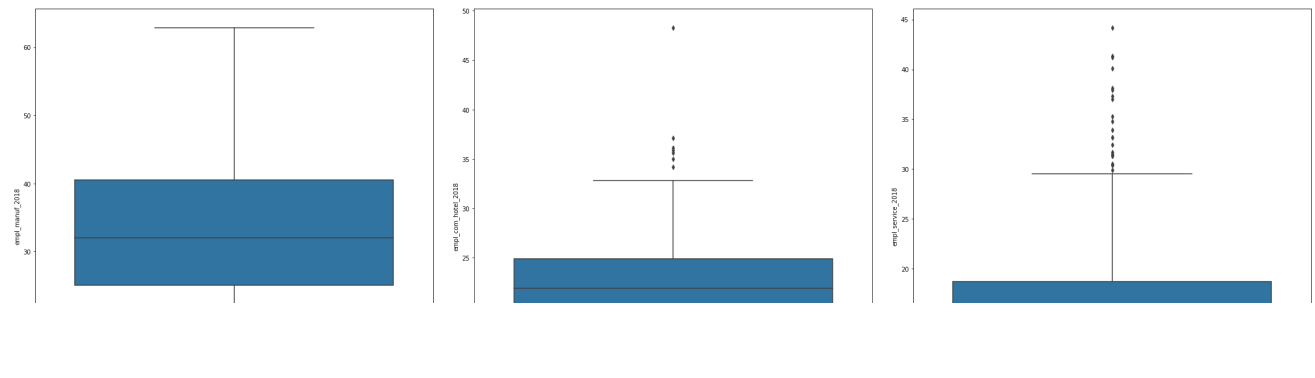

age_18_34_2017
age_35_59_2017
age4_60_74_2017
age_75_more_2017
disposable_inc_2016
gdp_2016
protection_total_2017
protection_open_2017
protection_accepted_2017
protection_rejected_2017
dwellings_now_2017
dwellings_2017
space_per_app_2017
space_per_inh_2017

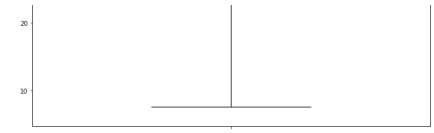

vehicles_2018
graduates_voc_2017
Absolventen/Abgänger allgemeinbildender Schulen 2017 - insgesamt ohne Externe (je 1000 Einwohner)

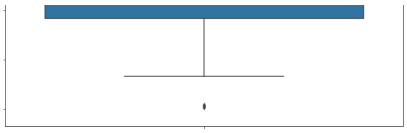

graduates_without_secondary_2017
graduates_lower_secondary_2017
graduates_secondary_2017

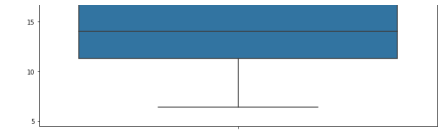

graduates_higher_2017
child_day_care_2018
business_reg_2017

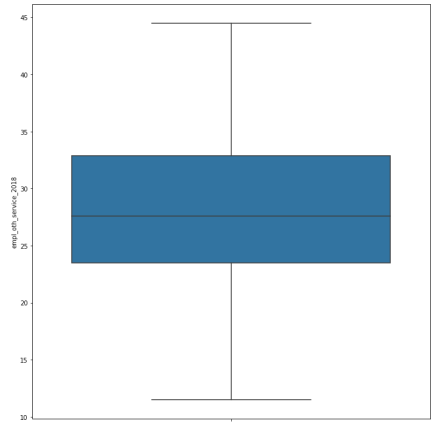

insolvencies_2017
empl_total_2018
empl_agr_2018

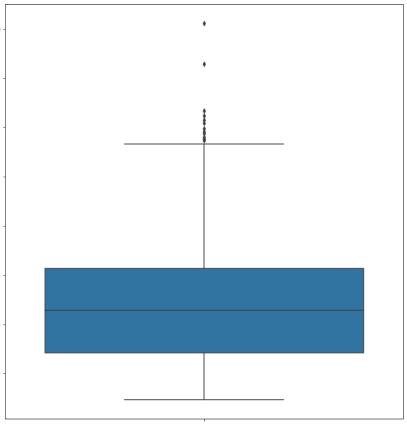

empl_manuf_2018
empl_com_hotel_2018
empl_service_2018

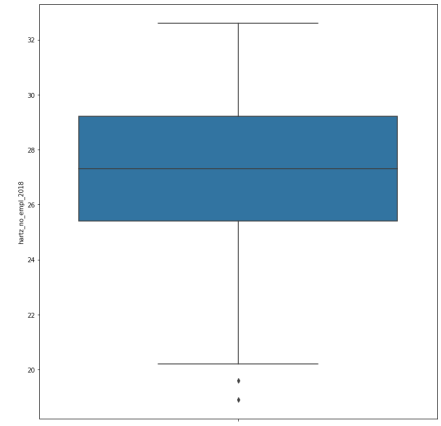


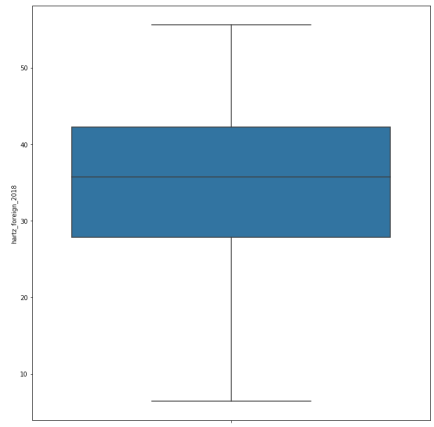


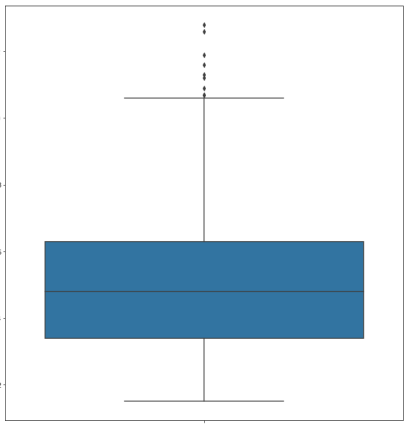


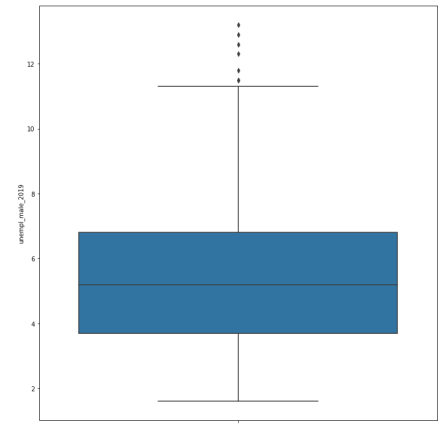

empl_oth_service_2018

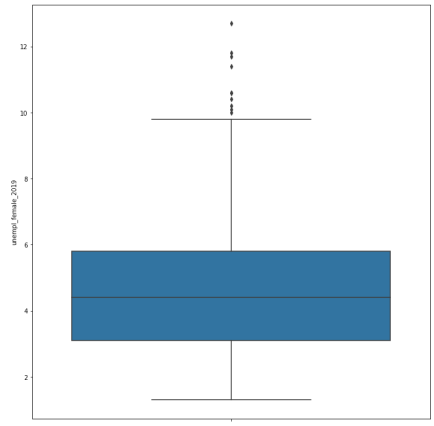

hartz_total_2018

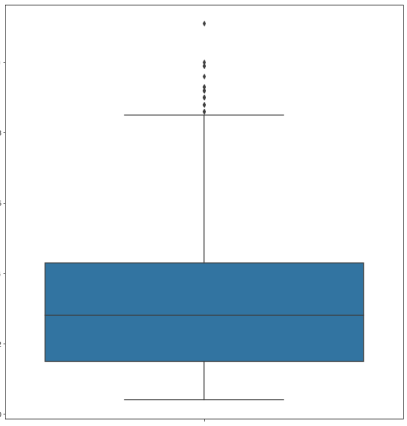

hartz_no_empl_2018

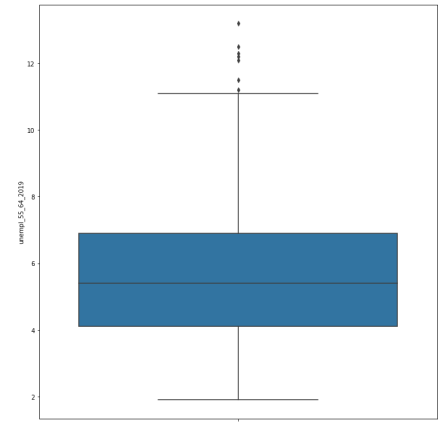

hartz_foreign_2018

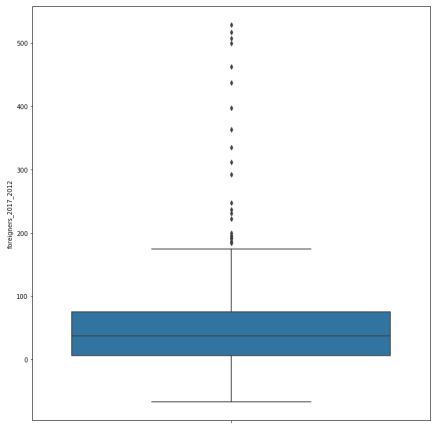

unempl_total_2019

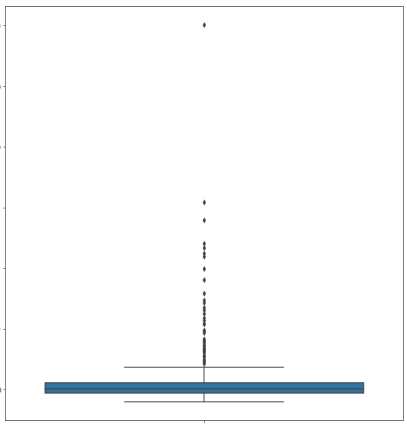

unempl_male_2019

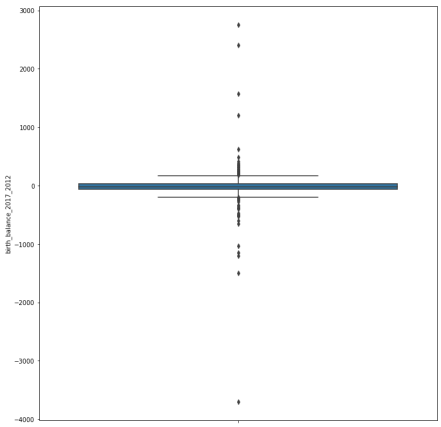

unempl_female_2019

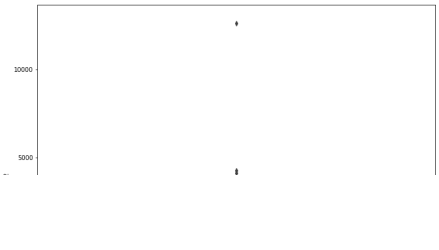

unempl_15_19_2019

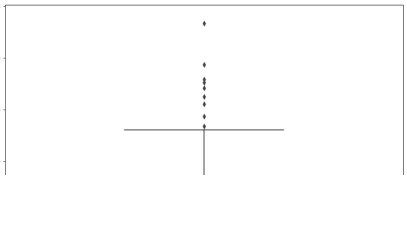

unempl_55_64_2019

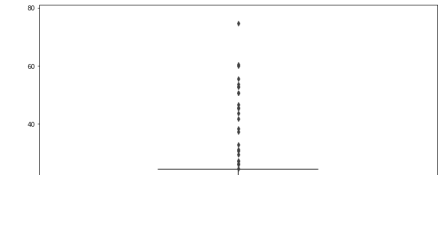

foreigners_2017_2012

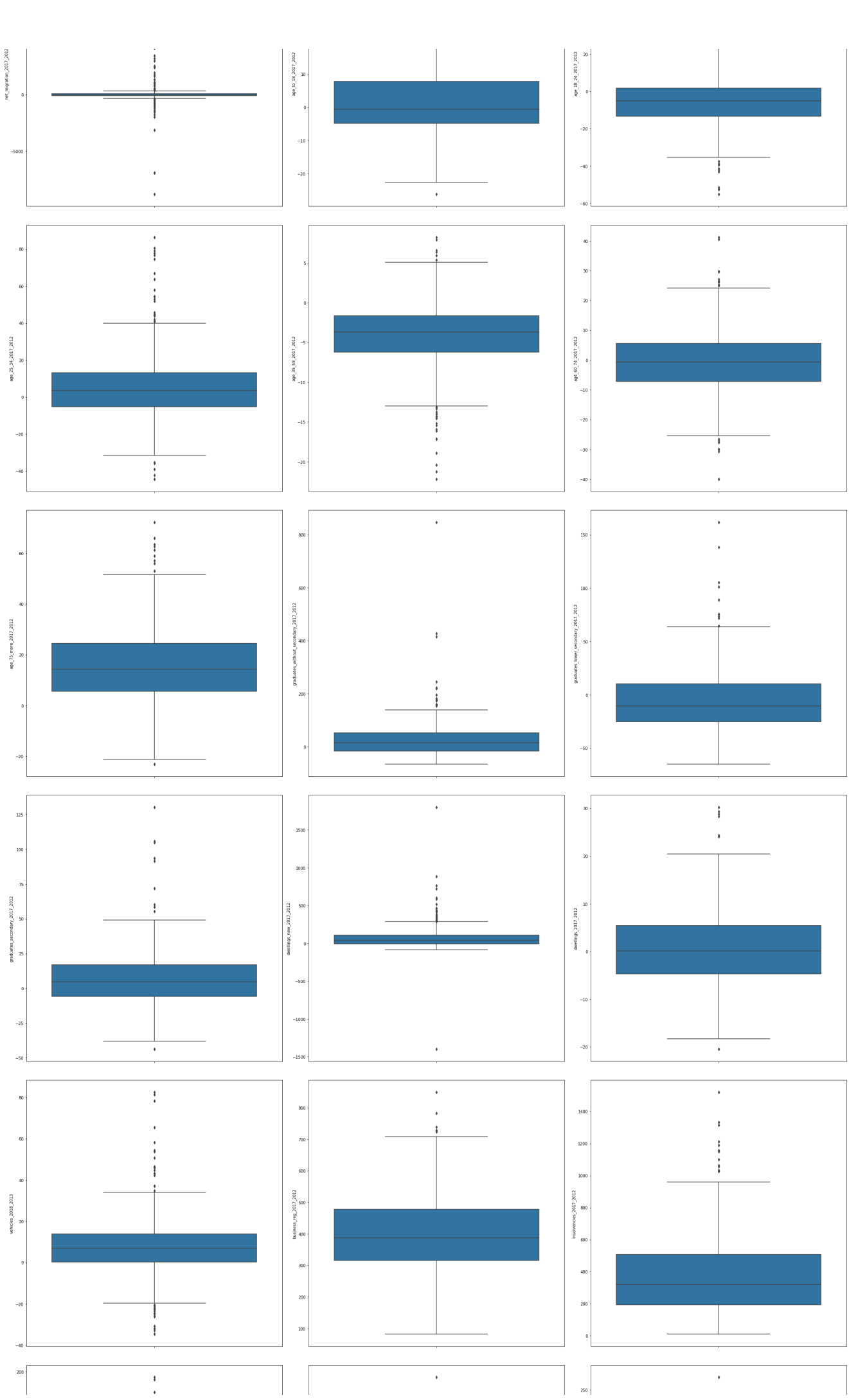

population_density_2017_2012

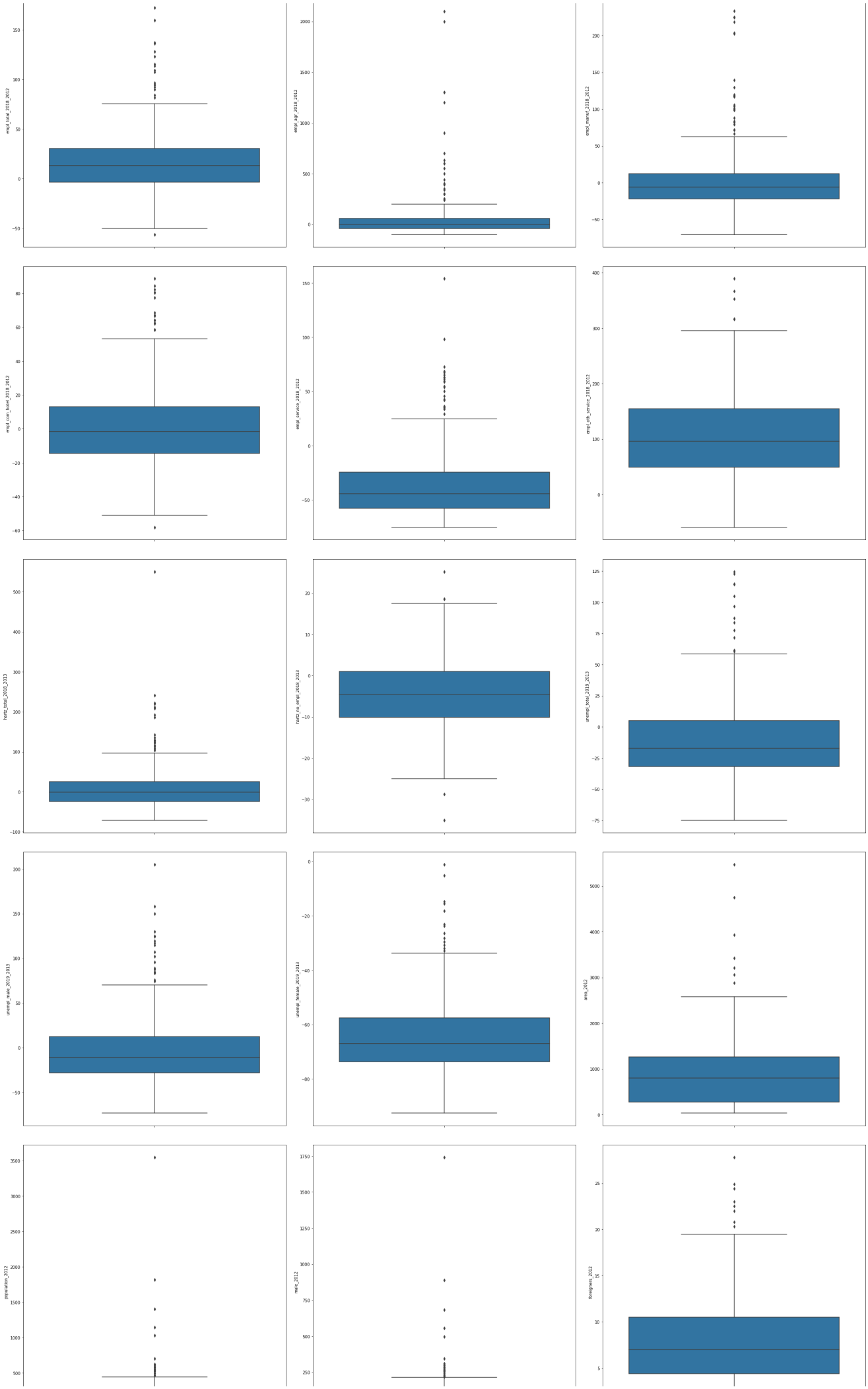

birth_balance_2017_2012

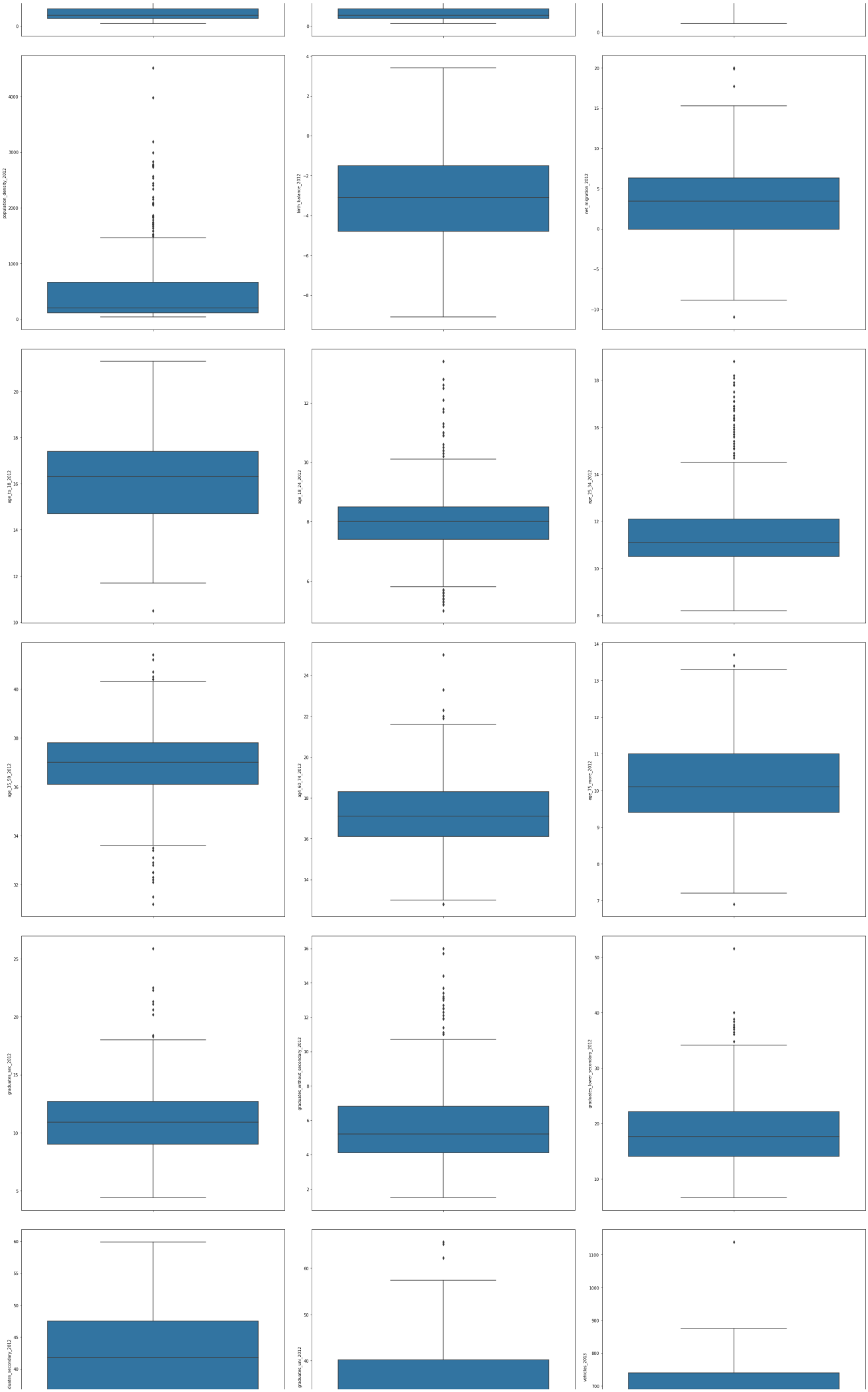


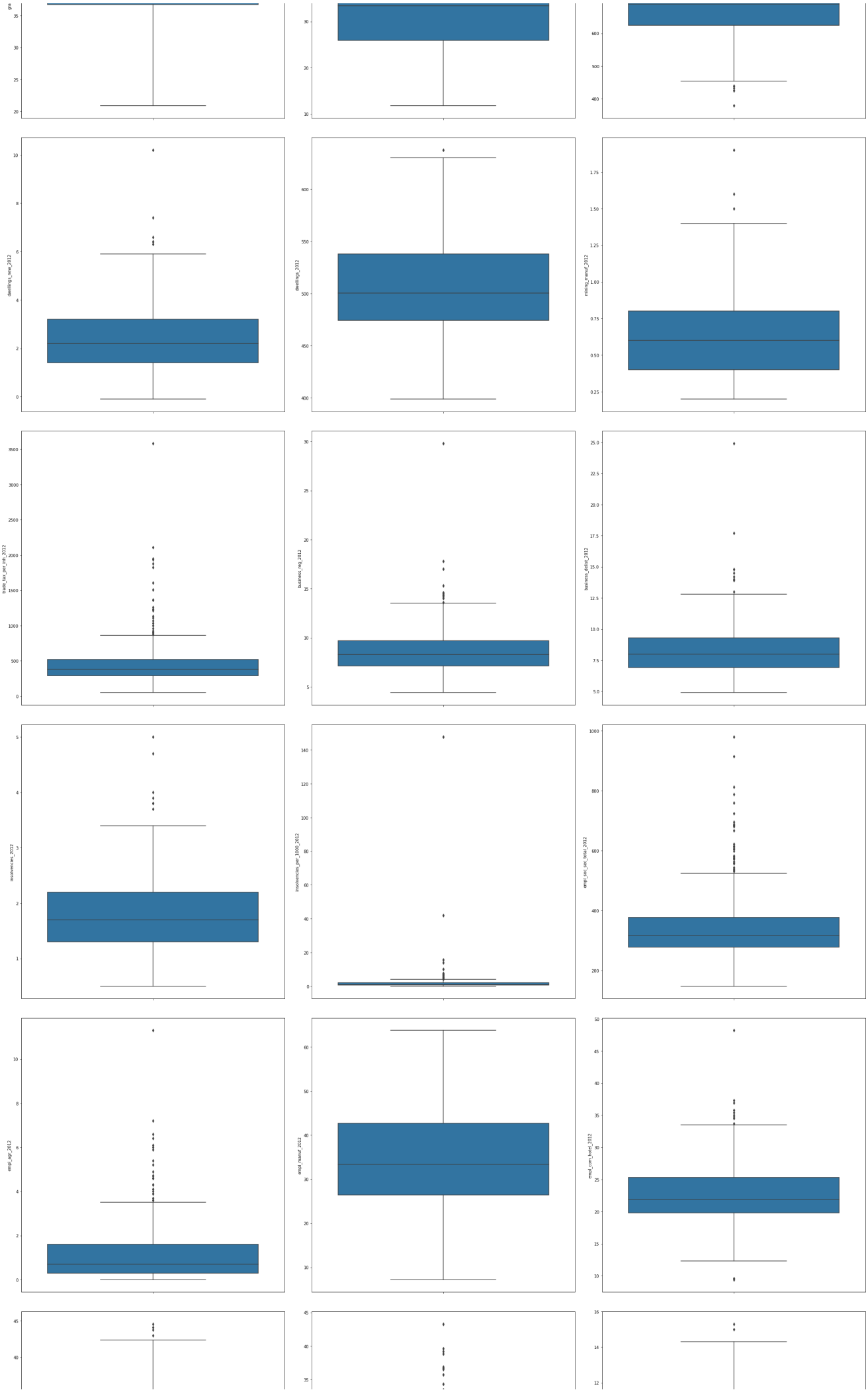


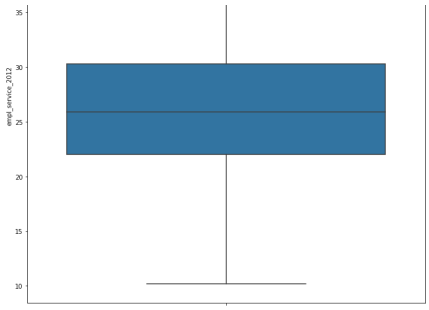


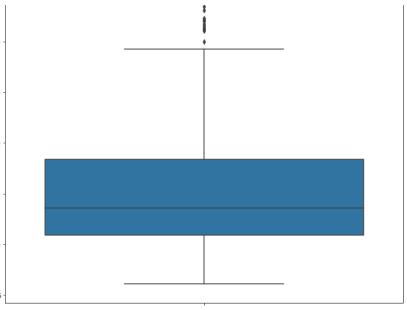

net_migration_2017_2012
age_to_18_2017_2012
age_18_24_2017_2012
age_25_34_2017_2012
age_35_59_2017_2012
age_60_74_2017_2012
age_75_more_2017_2012
graduates_without_secondary_2017_2012
graduates_lower_secondary_2017_2012
graduates_secondary_2017_2012
dwellings_new_2017_2012
dwellings_2017_2012
vehicles_2018_2013
business_reg_2017_2012
insolvencies_2017_2012

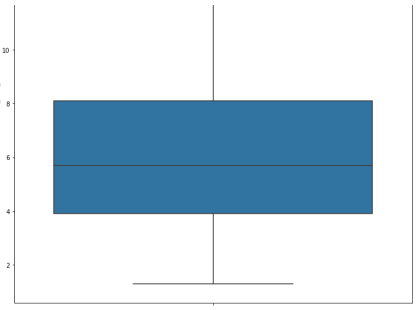

empl_total_2018_2012
empl_agr_2018_2012
empl_manuf_2018_2012
empl_com_hotel_2018_2012
empl_service_2018_2012
empl_oth_service_2018_2012
hartz_total_2018_2013
hartz_no_empl_2018_2013
unempl_total_2019_2013
unempl_male_2019_2013
unempl_female_2019_2013
area_2012
population_2012
male_2012
foreigners_2012

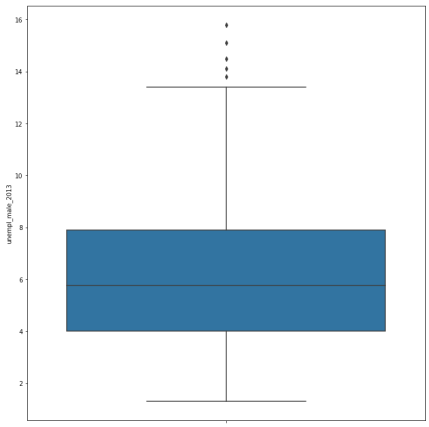

population_density_2012
birth_balance_2012
net_migration_2012
age_to_18_2012
age_18_24_2012
age_25_34_2012
age_35_59_2012
age_60_74_2012
age_75_more_2012
graduates_sec_2012
graduates_without_secondary_2012
graduates_lower_secondary_2012
graduates_secondary_2012
graduates_un_2012
vehicles_2013

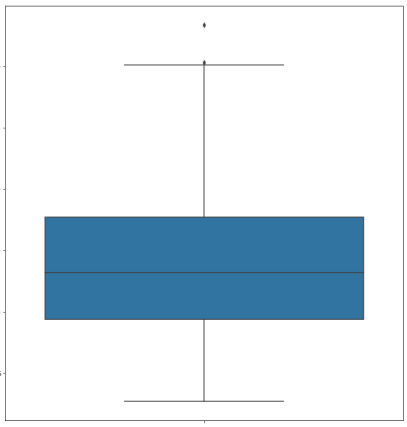

gfa
dwellings_new_2012
dwellings_2012
mining_manuf_2012
trade_tax_per_inh_2012
business_reg_2012
business_debit_2012
insolvencies_2012
insolvencies_per_1000_2012
empl_soc_sec_total_2012
empl_agr_2012
empl_manuf_2012
empl_com_hotel_2012

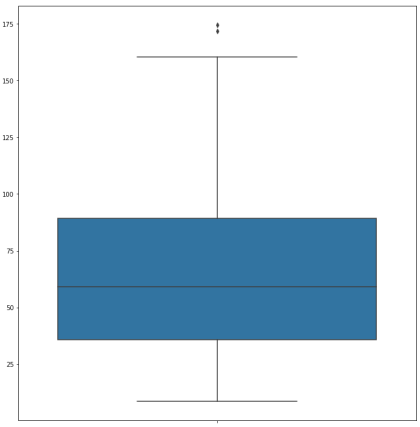

empl_service_2012

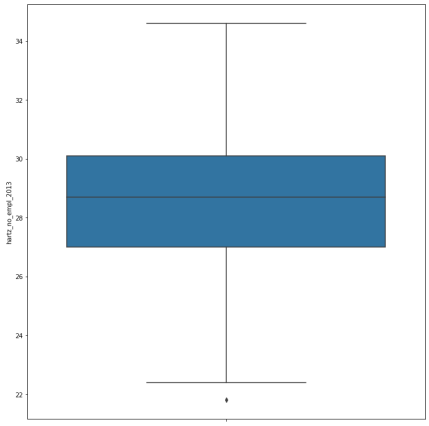

empl_oth_service_2012

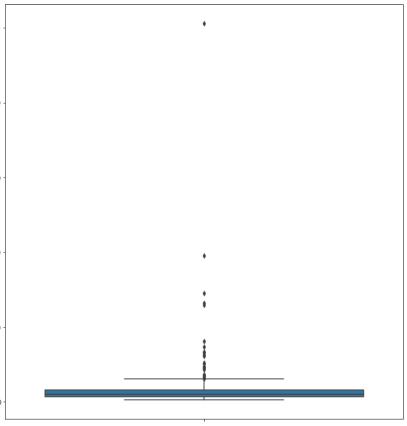

unempl_total_2013

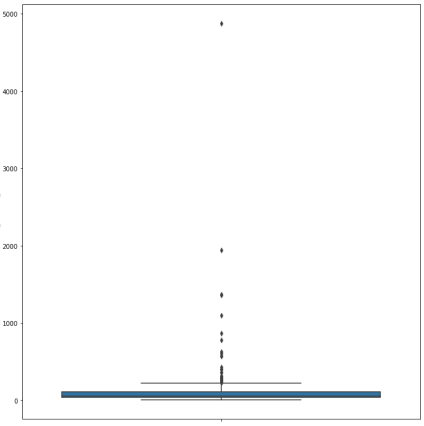

unempl_male_2013

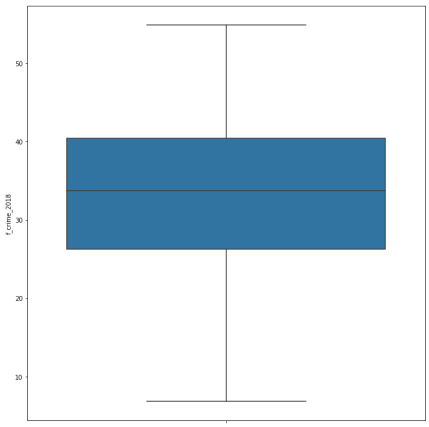

unempl_female_2013

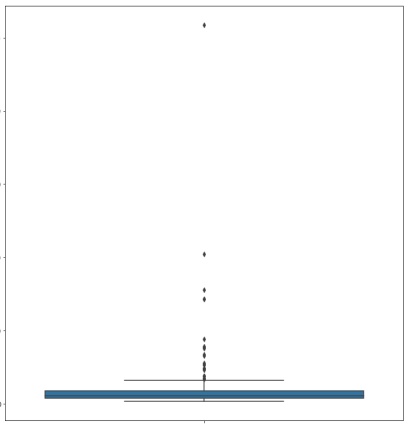

hartz_total_2013

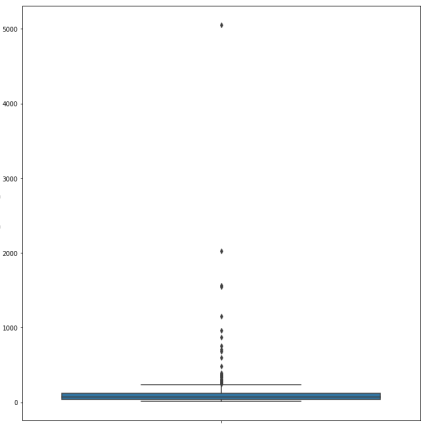

hartz_no_empl_2013

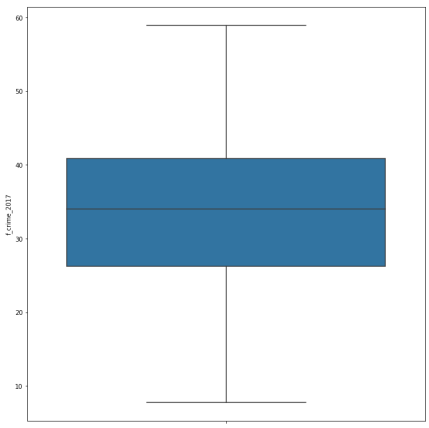

total_suspects_2018

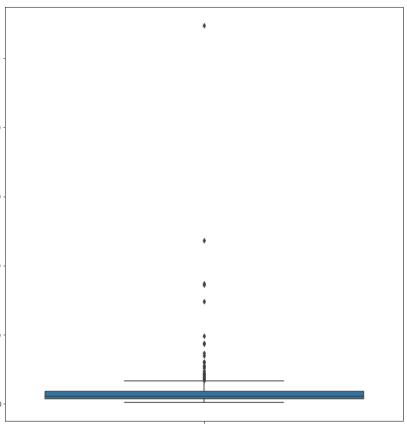

foreign_suspects_2018

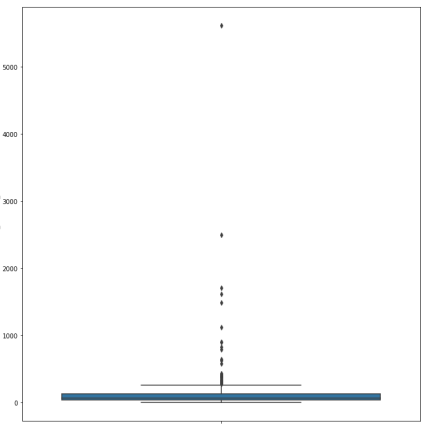

f_crime_2018

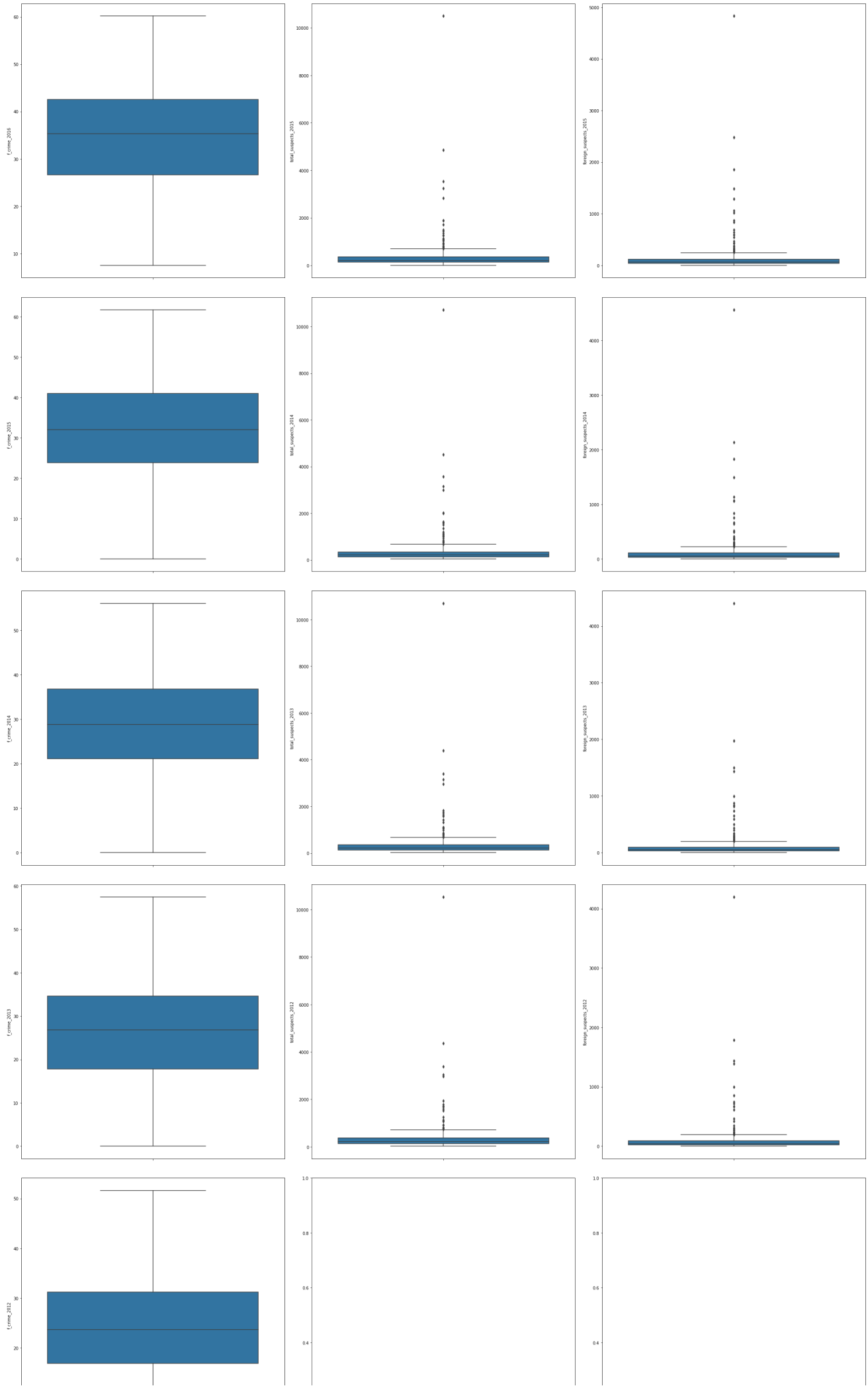

total_suspects_2017

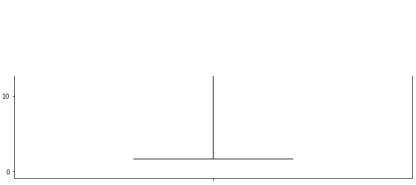
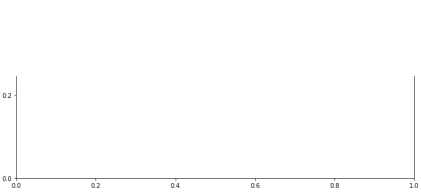
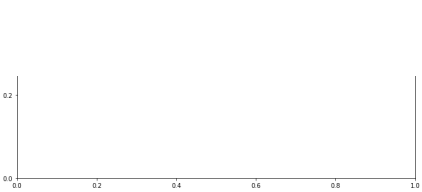
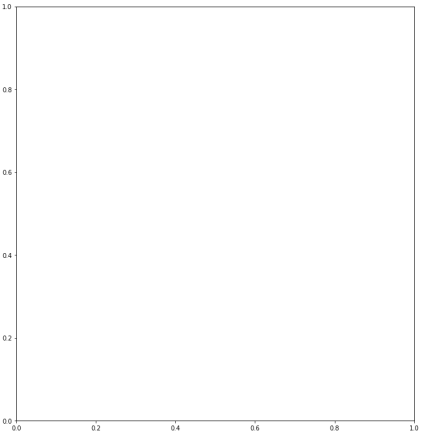
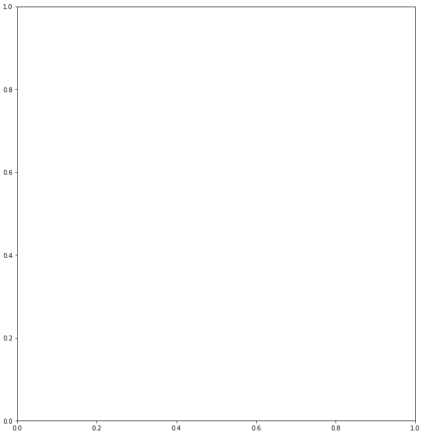
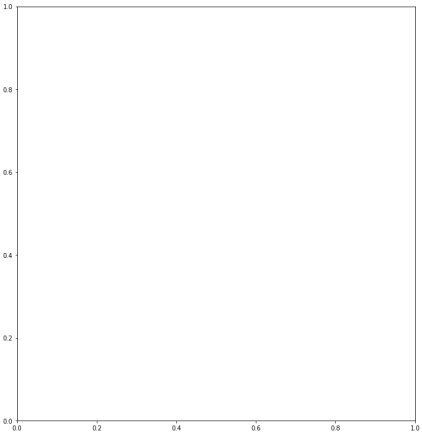
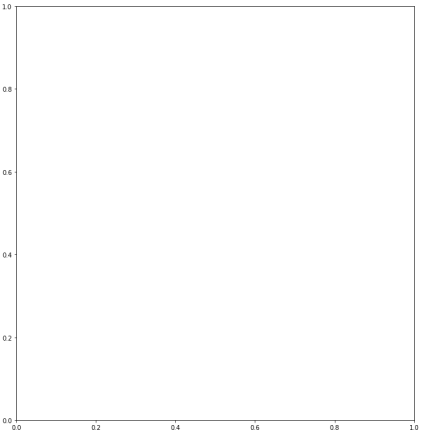
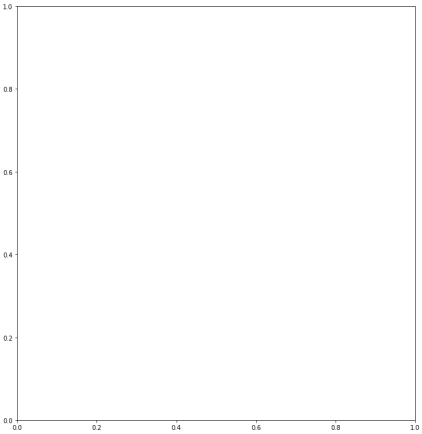
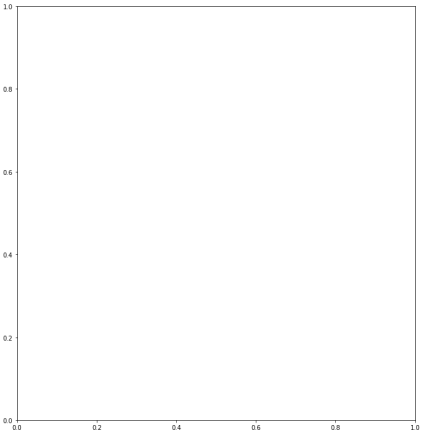
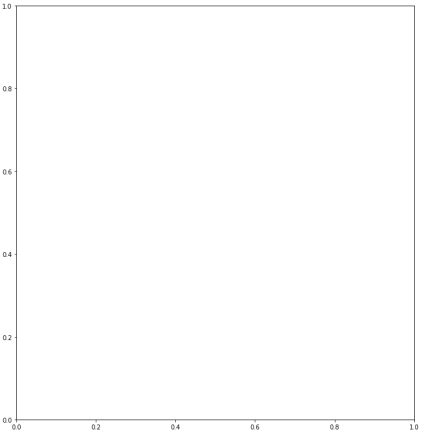
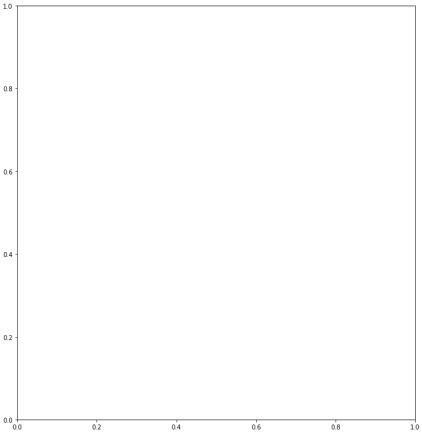
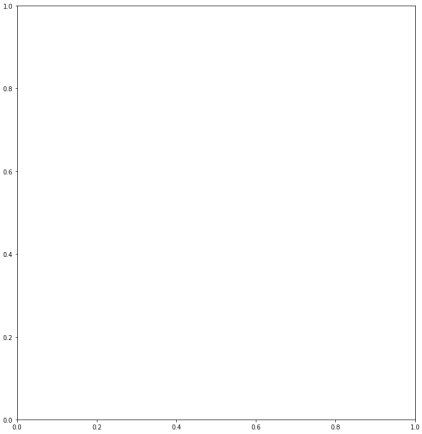
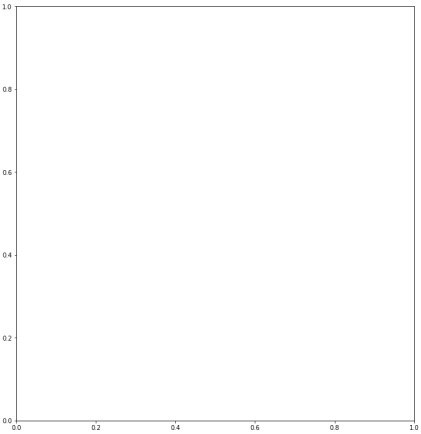
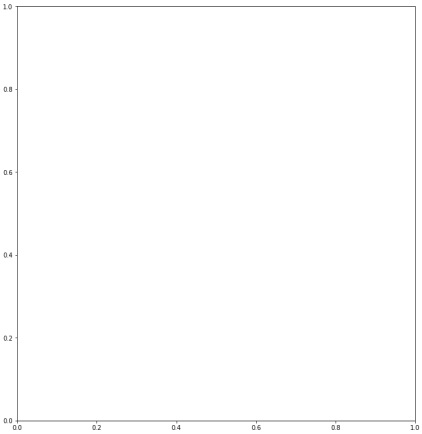
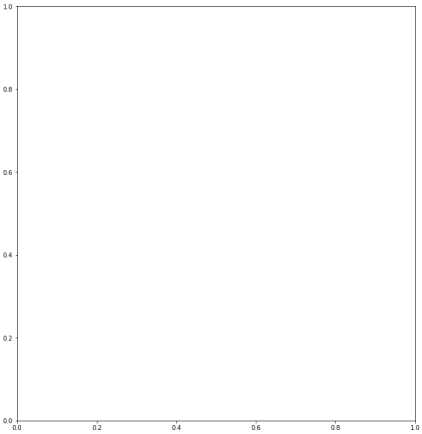
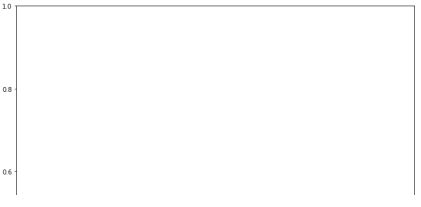
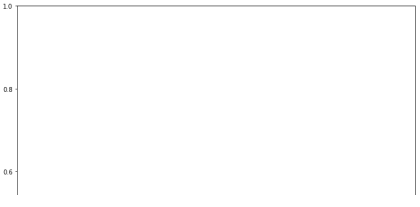
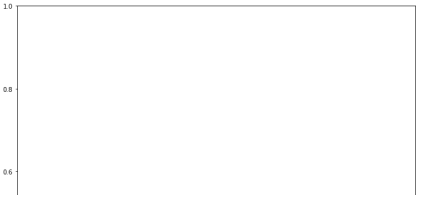

foreign_suspects_2017

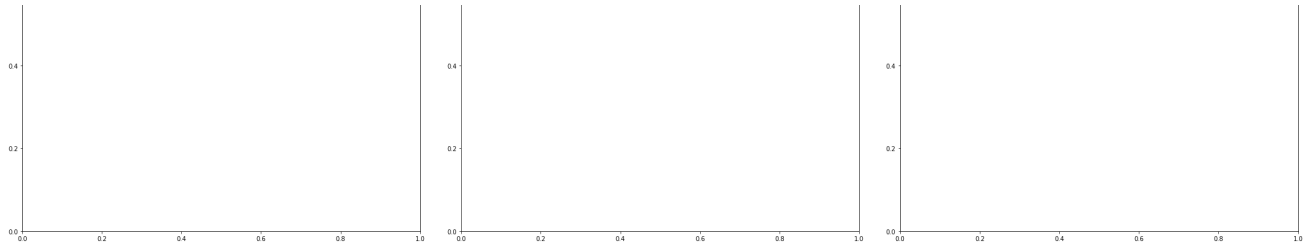

f_crime_2017


total_suspects_2016


foreign_suspects_2016


f_crime_2016
total_suspects_2015
foreign_suspects_2015
f_crime_2015
total_suspects_2014
foreign_suspects_2014
f_crime_2014
total_suspects_2013
foreign_suspects_2013
f_crime_2013
total_suspects_2012
foreign_suspects_2012
f_crime_2012



In [186]:

```
# we can even find them using the following code
def remove_outlier(df1):
    for i in df1.columns:
        q1 = df1[i].quantile(0.25)
        q3 = df1[i].quantile(0.75)
        iqr = q3-q1 #Interquartile range
        item1_low  = q1-1.5*iqr
        item1_high = q3+1.5*iqr
        df_out = df1.loc[(df1[i] < item1_low) | (df1[i] > item1_high)]
        return df_out
```

In [187]:

```
remove_outlier(df1)["vot19_14"] # the below are the ouliers in the data
```

Out[187]:

```
378     9.566501
379    18.453636
380     5.929222
381    10.577550
382     8.346655
383    14.983195
384    12.764963
385    16.782858
386    16.664091
387    16.973704
388    18.334998
389    15.582975
390    15.260989
391    17.965880
392    17.625163
393    15.364231
394    16.939578
395    19.574445
396    18.712021
397    13.389590
398    18.719615
399    15.881514
400    19.110053
Name: vot19_14, dtype: float64
```

feature selection and model building

model1 : feature selection using only features that are not highly colinear to each other

In [52]:

```
df3 = df1.copy()
df3.head(2)
```

Out[52]:

	Nr	subregion	vot19_14	turnout14	turnout19	turnout19_14	debt_2013	debt_2014	debt_2015	debt_2016	debt_2017	debt_2018

	Nr	subregion	vot19_14	turnout_14	turnout_18	turnout_19	debt_2013	debt_2014	debt_2015	debt_2016	debt_2017	debt_2018
0	1001	1.0	0.003027	0.357401	0.502829	0.209524	12.11	12.10	12.24	12.30	12.23	12.03
1	1002	1.0	0.060316	0.402589	0.588603	0.186015	12.04	12.03	12.17	12.23	12.16	11.96

In [53]:

```
from sklearn.model_selection import train_test_split
X= df3.drop("vot19_14", axis = 1)
Y = df3["vot19_14"]
```

In [54]:

```
corr = X.corr()
columns = np.full((corr.shape[0],), True, dtype=bool)
for i in range(corr.shape[0]):
    for j in range(i+1, corr.shape[0]):
        if corr.iloc[i,j] >= 0.8:
            if columns[j]:
                columns[j] = False
```

In [55]:

```
corelated_columns = X.columns[columns]
corelated_columns.shape     # these are highly correlated features which have to be removed
```

Out[55]:

```
(78,)
```

In [56]:

```
X.drop(X.columns[columns], axis = 1, inplace = True)
```

In [53]:

```
### moded1 : using stats model why eliminating the correlated features from the data frame


x1 = sm.add_constant(X) #Adding the constant
model = sm.OLS(Y,x1).fit() # fitting the model
print(model.summary()) # model summary


def checkVIF(x):
    vif = pd.DataFrame()
    vif['Features'] = x.columns
    vif['VIF'] = [variance_inflation_factor(x.values, i) for i in range(x.shape[1])]
    vif['VIF'] = round(vif['VIF'], 2)
    vif = vif.sort_values(by = "VIF", ascending = False)
    return(vif)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                vot19_14   R-squared:                       0.902
Model:                            OLS   Adj. R-squared:                  0.882
Method:                 Least Squares   F-statistic:                     44.32
Date:                Wed, 29 Jan 2020   Prob (F-statistic):          2.30e-131
Time:                        00:25:45   Log-Likelihood:                -771.13
No. Observations:                 401   AIC:                             1682.
Df Residuals:                     331   BIC:                             1962.
Df Model:                          69
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                        117.5092    209.038      0.562      0.574    -293.701      528.719
subregion                      0.4018      0.060      6.664      0.000       0.283        0.520
debt_2014                     -0.9748      0.654     -1.491      0.137      -2.261        0.311
debt_2015                     -0.0712      0.817     -0.087      0.931      -1.679        1.536
```

debt_2015	0.0712	0.617	0.115	0.908	-1.079	1.222
debt_2016	0.6394	1.171	0.546	0.585	-1.664	2.943
debt_2017	-0.9414	1.373	-0.686	0.493	-3.643	1.760
debt_2018	1.4874	0.881	1.688	0.092	-0.246	3.221
germans_2017	-0.0123	0.020	-0.627	0.531	-0.051	0.026
age_75_more_2017	0.1428	0.355	0.402	0.688	-0.556	0.842
space_per_inh_2017	-0.0653	0.050	-1.318	0.188	-0.163	0.032
empl_total_2018	-0.0025	0.008	-0.332	0.740	-0.017	0.012
hartz_total_2018	0.0255	0.019	1.319	0.188	-0.013	0.063
unempl_total_2019	2.3458	2.868	0.818	0.414	-3.295	7.987
unempl_male_2019	-2.0287	1.590	-1.276	0.203	-5.156	1.099
unempl_female_2019	-0.5468	1.343	-0.407	0.684	-3.188	2.094
unempl_15_19_2019	-0.1701	0.125	-1.366	0.173	-0.415	0.075
unempl_55_64_2019	0.7541	0.213	3.543	0.000	0.335	1.173
age_25_34_2017_2012	-0.0051	0.008	-0.643	0.520	-0.021	0.011
empl_service_2018_2012	0.0077	0.006	1.225	0.222	-0.005	0.020
unempl_total_2019_2013	-0.0286	0.019	-1.474	0.141	-0.067	0.010
unempl_male_2019_2013	0.0242	0.017	1.423	0.156	-0.009	0.058
area_2012	-0.0005	0.000	-1.821	0.069	-0.001	3.79e-05
population_2012	-0.0497	0.049	-1.006	0.315	-0.147	0.048
male_2012	0.1316	0.091	1.440	0.151	-0.048	0.311
foreigners_2012	-0.2712	0.072	-3.758	0.000	-0.413	-0.129
population_density_2012	-0.0011	0.000	-2.626	0.009	-0.002	-0.000
birth_balance_2012	-0.1974	0.164	-1.206	0.229	-0.519	0.125
age_to_18_2012	-0.5630	1.423	-0.396	0.693	-3.363	2.237
age_18_24_2012	-1.5267	1.414	-1.080	0.281	-4.307	1.254
age_25_34_2012	0.4041	1.436	0.281	0.779	-2.421	3.229
age_35_59_2012	-0.2232	1.410	-0.158	0.874	-2.997	2.550
ag4_60_74_2012	-0.7916	1.410	-0.561	0.575	-3.566	1.982
age_75_more_2012	-0.4221	1.429	-0.295	0.768	-3.233	2.389
graduates_sec_2012	-0.1359	0.067	-2.029	0.043	-0.268	-0.004
graduates_lower_secondary_2012	-0.1273	0.026	-4.942	0.000	-0.178	-0.077
graduates_uni_2012	-0.0179	0.019	-0.942	0.347	-0.055	0.019
vehicles_2013	0.0010	0.003	0.360	0.719	-0.004	0.006
dwellings_2012	-0.0024	0.006	-0.379	0.705	-0.015	0.010
trade_tax_per_inh_2012	0.0002	0.001	0.286	0.775	-0.001	0.001
business_delist_2012	-0.1641	0.076	-2.152	0.032	-0.314	-0.014
empl_soc_sec_total_2012	-0.0006	0.008	-0.069	0.945	-0.017	0.015
empl_agr_2012	-0.5554	1.544	-0.360	0.719	-3.593	2.483
empl_manuf_2012	-0.6388	1.539	-0.415	0.678	-3.667	2.389
empl_com_hotel_2012	-0.7250	1.538	-0.471	0.638	-3.751	2.301
empl_service_2012	-0.6750	1.539	-0.438	0.661	-3.703	2.353
empl_oth_service_2012	-0.7842	1.543	-0.508	0.612	-3.819	2.250
unempl_total_2013	-1.1879	0.500	-2.374	0.018	-2.172	-0.204
unempl_male_2013	1.1205	0.468	2.393	0.017	0.200	2.041
unempl_female_2013	-0.0073	0.046	-0.158	0.874	-0.098	0.084
hartz_total_2013	0.0259	0.022	1.198	0.232	-0.017	0.069
total_suspects_2018	-0.0028	0.004	-0.682	0.495	-0.011	0.005
foreign_suspects_2018	0.0126	0.009	1.457	0.146	-0.004	0.030
total_suspects_2017	0.0010	0.005	0.187	0.852	-0.009	0.011
foreign_suspects_2017	-0.0082	0.011	-0.750	0.454	-0.030	0.013
f_crime_2017	0.0123	0.028	0.440	0.660	-0.043	0.067
total_suspects_2016	-0.0038	0.005	-0.804	0.422	-0.013	0.006
foreign_suspects_2016	-0.0007	0.009	-0.075	0.940	-0.019	0.017
f_crime_2016	-0.0168	0.028	-0.596	0.551	-0.072	0.039
total_suspects_2015	-0.0008	0.005	-0.150	0.881	-0.011	0.009
foreign_suspects_2015	0.0092	0.010	0.913	0.362	-0.011	0.029
f_crime_2015	0.0025	0.026	0.096	0.924	-0.049	0.054
total_suspects_2014	-0.0015	0.003	-0.560	0.576	-0.007	0.004
foreign_suspects_2014	0.0012	0.008	0.149	0.882	-0.015	0.017
f_crime_2014	0.0101	0.025	0.407	0.685	-0.039	0.059
total_suspects_2013	0.0033	0.004	0.806	0.421	-0.005	0.011
foreign_suspects_2013	-0.0111	0.010	-1.145	0.253	-0.030	0.008
f_crime_2013	0.0071	0.026	0.275	0.784	-0.044	0.058
total_suspects_2012	0.0020	0.004	0.476	0.635	-0.006	0.010
foreign_suspects_2012	-0.0023	0.011	-0.216	0.829	-0.023	0.019
f_crime_2012	-0.0159	0.030	-0.529	0.597	-0.075	0.043

Omnibus:	11.959	Durbin-Watson:	1.721
Prob(Omnibus):	0.003	Jarque-Bera (JB):	15.978
Skew:	0.266	Prob(JB):	0.000339
Kurtosis:	3.821	Cond. No.	5.49e+06

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.49e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

strong multicollinearity or other numerical problems.

In [56]:

```
from scipy.stats import shapiro
shapiro(model.resid)
```

Out[56]:

```
(0.9914925694465637, 0.021149726584553 72)
```

From the Shapiro results, since pvalue is less than 0.05 we can conclude that residuals do not follow normal distribution

cooks distance

In [54]:

```
influence = model.get_influence()
cook = influence.summary_frame()
cook
```

Out[54]:

	dfb_const	dfb_subregion	dfb_debt_2014	dfb_debt_2015	dfb_debt_2016	dfb_debt_2017	dfb_debt_2018	dfb_germans_2017	dfb_a
0	0.042052	-0.038832	0.020715	0.012420	-0.051741	0.038788	-0.011636	-0.013273	
1	0.006215	0.026378	0.016933	-0.006022	-0.010303	0.001892	0.006012	-0.003712	
2	-0.032482	-0.067706	0.027350	-0.032855	0.010411	0.030986	-0.051268	-0.016464	
3	0.005221	-0.006651	-0.002768	0.002048	-0.009344	-0.001411	0.016436	-0.008095	
4	-0.057230	-0.064530	0.082622	-0.110199	-0.007166	0.023843	0.013816	-0.028787	
...	...	...	...	...	...	...	...	...	
396	0.002100	0.042552	-0.051500	0.015809	0.027206	-0.029299	0.016533	0.018497	
397	0.000327	0.000063	-0.000673	0.000623	-0.000293	0.000361	-0.000194	0.000402	
398	0.095855	0.119944	0.163999	-0.088456	-0.037784	0.015610	-0.018229	-0.100427	
399	0.011189	-0.006101	0.000047	-0.000768	0.013067	-0.010805	-0.002243	0.007272	
400	-0.118078	0.049424	0.040498	-0.027325	0.013209	-0.045417	0.024348	-0.068116	

401 rows × 76 columns

In [55]:

```
cook[cook.cooks_d > 0.1]
```

Out[55]:

	dfb_const	dfb_subregion	dfb_debt_2014	dfb_debt_2015	dfb_debt_2016	dfb_debt_2017	dfb_debt_2018	dfb_germans_2017	dfb_a
66	0.016153	-0.146661	-0.085222	0.016375	0.144280	-0.132451	0.036901	0.084133	
324	-0.033384	0.130111	0.001108	0.010598	-0.025763	-0.007164	0.023979	0.409991	
395	-0.051812	-0.144855	0.521765	-0.221503	-0.303196	0.063467	0.184109	-0.017223	

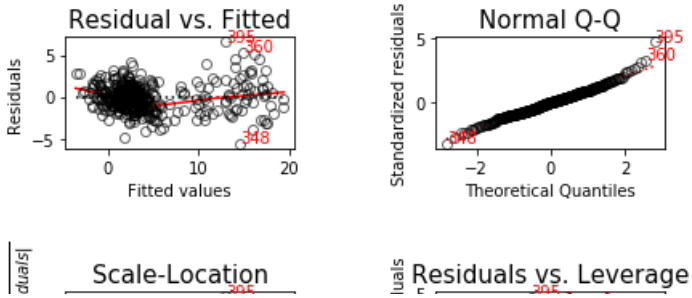
In [58]:

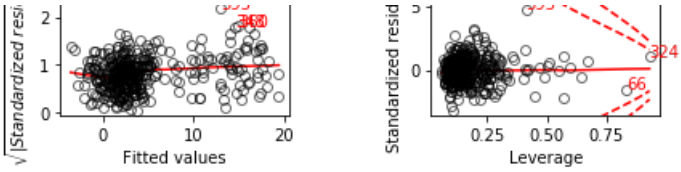
```
!pip install lmdiag
```

```
Collecting lmdiag
  Downloading
https://files.pythonhosted.org/packages/db/76/8b15c0e5065156fa776fe117877a18ff91f2218fb35d79d70b9b5
a59/lmdiag-0.3.7-py3-none-any.whl
Collecting linearmodels (from lmdiag)
  Downloading
https://files.pythonhosted.org/packages/dd/3f/ab18e5b2dc88be8e78974c5518813b0b53785a76cffa8928e4d62
cc9/linearmodels-4.16-cp37-cp37m-win_amd64.whl (1.6MB)
Requirement already satisfied: scipy in c:\users\sourav ch\anaconda3\lib\site-packages (from
lmdiag) (1.3.1)
Requirement already satisfied: numpy in c:\users\sourav ch\anaconda3\lib\site-packages (from
lmdiag) (1.16.5)
Requirement already satisfied: statsmodels in c:\users\sourav ch\anaconda3\lib\site-packages (from
lmdiag) (0.10.1)
Requirement already satisfied: matplotlib in c:\users\sourav ch\anaconda3\lib\site-packages (from
lmdiag) (3.1.1)
Requirement already satisfied: pandas in c:\users\sourav ch\anaconda3\lib\site-packages (from
lmdiag) (0.25.1)
Collecting property-cached>=1.6.3 (from linearmodels->lmdiag)
  Downloading
https://files.pythonhosted.org/packages/c0/11/6e91ff5fe0476492f023cebad434a1a34fc513cfa98ddb1f3e5c8
d99/property_cached-1.6.3-py2.py3-none-any.whl
Collecting mypy-extensions>=0.4 (from linearmodels->lmdiag)
  Downloading
https://files.pythonhosted.org/packages/5c/eb/975c7c080f3223a5cdaff09612f3a5221e4ba534f7039db34c35c
6a5/mypy_extensions-0.4.3-py2.py3-none-any.whl
Collecting Cython>=0.29.14 (from linearmodels->lmdiag)
  Downloading
https://files.pythonhosted.org/packages/1f/be/b14be5c3ad1ff73096b518be1538282f053ec34faaca60a8753d9
e93/Cython-0.29.14-cp37-cp37m-win_amd64.whl (1.7MB)
Requirement already satisfied: patsy in c:\users\sourav ch\anaconda3\lib\site-packages (from
linearmodels->lmdiag) (0.5.1)
Requirement already satisfied: cycler>=0.10 in c:\users\sourav ch\anaconda3\lib\site-packages
(from matplotlib->lmdiag) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\sourav ch\anaconda3\lib\site-packages
(from matplotlib->lmdiag) (1.1.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in c:\users\sourav
ch\anaconda3\lib\site-packages (from matplotlib->lmdiag) (2.4.2)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\sourav ch\anaconda3\lib\site-
packages (from matplotlib->lmdiag) (2.8.0)
Requirement already satisfied: pytz>=2017.2 in c:\users\sourav ch\anaconda3\lib\site-packages
(from pandas->lmdiag) (2019.3)
Requirement already satisfied: six in c:\users\sourav ch\anaconda3\lib\site-packages (from patsy->
linearmodels->lmdiag) (1.12.0)
Requirement already satisfied: setuptools in c:\users\sourav ch\anaconda3\lib\site-packages (from
kiwisolver>=1.0.1->matplotlib->lmdiag) (41.4.0)
Installing collected packages: property-cached, mypy-extensions, Cython, linearmodels, lmdiag
  Found existing installation: Cython 0.29.13
    Uninstalling Cython-0.29.13:
      Successfully uninstalled Cython-0.29.13
Successfully installed Cython-0.29.14 linearmodels-4.16 lmdiag-0.3.7 mypy-extensions-0.4.3
property-cached-1.6.3
```

In [59]:

```
import lmdiag
import matplotlib.pyplot as plt
lmdiag.plot(model)
plt.show()
```





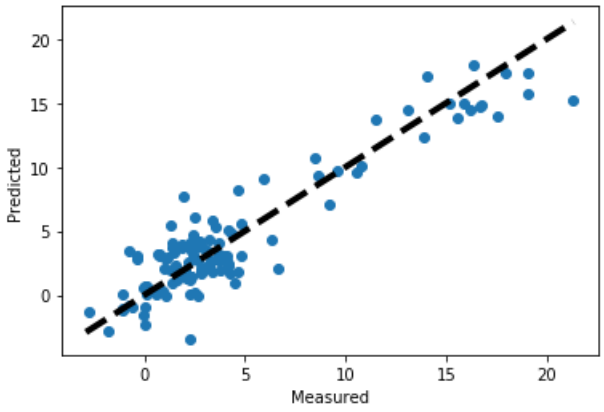
In [263]:

```
####model 1: using scikit learn to predict the MSe metric for model evaluation



X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.3, random_state = 10)
reg = LinearRegression()
m1 = reg.fit(X_train, Y_train)
#Predict the model on train
pred_train = m1.predict(X_train)
print("MSE on train: ", mean_squared_error(Y_train,pred_train))
print("RSquared: ",m1.score(X_train,Y_train))
#Predict on test
pred_test = m1.predict(X_test)
print("MSE on test: ", mean_squared_error(Y_test,pred_test))
print("RSquared: ",m1.score(X_test,Y_test))
```

```
MSE on train:  2.7102603797808373
RSquared:  0.9048023720966871
MSE on test:  3.9235577435999858
RSquared:  0.8552997473777733
```

In [264]:

```
fig, ax = plt.subplots()
ax.scatter(Y_test, pred_test)
ax.plot([Y.min(), Y.max()], [Y.min(), Y.max()], 'k--', lw=4)
ax.set_xlabel('Measured')
ax.set_ylabel('Predicted')
plt.show()
```



In [24]:

```
df4 = df1.copy()
```

In [23]:

```
df4.drop(columns = {'subregion',"Nr"}, inplace = True)
```

model2 : using p-value backward elimination we have the following model and scores

In [25]:

```
x = df4.drop("vot19_14", axis = 1)
Y = df4["vot19_14"]




cols = list(x.columns)
pmax = 1
while (len(cols)>0):
    p= []
    X_1 = x[cols]
    X_1 = sm.add_constant(X_1)
    model2 = sm.OLS(Y,X_1).fit()
    p = pd.Series(model2.pvalues.values[1:],index = cols)
    pmax = max(p)
    print('pmsx:',pmax)
    feature_with_p_max = p.idxmax()
    print('feature_pmax',feature_with_p_max)
    if(pmax>0.05):
        cols.remove(feature_with_p_max)

    else:
        break
selected_features_BE = cols
print(selected_features_BE)
```

```
C:\Users\SOURAV CH\Anaconda3\lib\site-packages\numpy\core\fromnumeric.py:2389: FutureWarning: Meth
od .ptp is deprecated and will be removed in a future version. Use numpy.ptp instead.
  return ptp(axis=axis, out=out, **kwargs)
```

```
pmsx: 0.9823014416785996
feature_pmax insolvencies_2012
pmsx: 0.9751402790586139
feature_pmax insolvencies_per_1000_2012
pmsx: 0.976312864213492
feature_pmax business_reg_2017_2012
pmsx: 0.974295537685532
feature_pmax debt_2014
pmsx: 0.9679103543476841
feature_pmax turnout19
pmsx: 0.9470156728153247
feature_pmax age_25_34_2012
pmsx: 0.9457399426133217
feature_pmax empl_total_2018_2012
pmsx: 0.9479636414566406
feature_pmax empl_manuf_2018_2012
pmsx: 0.949584695398314
feature_pmax foreign_suspects_2016
pmsx: 0.9433821248657293
feature_pmax debt_2018
pmsx: 0.9343899013975979
feature_pmax ag4_60_74_2017
pmsx: 0.9293552031382638
feature_pmax debt_2015
pmsx: 0.9058374210164626
feature_pmax empl_oth_service_2018
pmsx: 0.8878620367978666
feature_pmax disposable_inc_2016
pmsx: 0.8667767907475088
feature_pmax foreigners_2012
pmsx: 0.8619620449332492
feature_pmax empl_agr_2012
pmsx: 0.8423788710464833
feature_pmax business_reg_2012
pmsx: 0.9031971761852674
feature_pmax business_delist_2012
pmsx: 0.8538770117932553
feature_pmax debt_2017
pmsx: 0.9300477480470293
feature_pmax debt_2016
pmsx: 0.8287923441805888
feature_pmax net_migration_2017_2012
pmsx: 0.832103871485715
feature_pmax unempl_female_2013
pmsx: 0.8129612471043786
feature_pmax hartz_total_2013
```

feature_pmax [illegible]
pmsx: 0.7984813668959626
feature_pmax net_migration_2012
pmsx: 0.7904412428985786
feature_pmax total_suspects_2015
pmsx: 0.8659751055058942
feature_pmax f_crime_2015
pmsx: 0.7689382985107848
feature_pmax foreign_suspects_2018
pmsx: 0.8088518948240919
feature_pmax total_suspects_2018
pmsx: 0.7844737903753496
feature_pmax unempl_15_19_2019
pmsx: 0.7769804638934206
feature_pmax foreigners_2017_2012
pmsx: 0.6987175442239892
feature_pmax birth_balance_2017
pmsx: 0.7385590891284994
feature_pmax graduates_sec_2012
pmsx: 0.7263427024913935
feature_pmax foreigners_2017
pmsx: 0.6635296388411157
feature_pmax population_density_2017_2012
pmsx: 0.655230343958111
feature_pmax dwellings_new_2017
pmsx: 0.6845138717746047
feature_pmax empl_soc_sec_total_2012
pmsx: 0.8019538096050057
feature_pmax empl_total_2018
pmsx: 0.6316625253034049
feature_pmax age_35_59_2017_2012
pmsx: 0.6286762818896834
feature_pmax age_25_34_2017_2012
pmsx: 0.6094106552727401
feature_pmax total_suspects_2013
pmsx: 0.5875363663101183
feature_pmax area_2012
pmsx: 0.5394363139836856
feature_pmax f_crime_2016
pmsx: 0.5546190522982204
feature_pmax population_2012
pmsx: 0.46283500926569354
feature_pmax hartz_no_empl_2018_2013
pmsx: 0.5036827937753559
feature_pmax hartz_no_empl_2018
pmsx: 0.5174393015822581
feature_pmax f_crime_2014
pmsx: 0.47976281733596504
feature_pmax protection_rejected_2017
pmsx: 0.911631503572509
feature_pmax protection_accepted_2017
pmsx: 0.6240419233086596
feature_pmax protection_open_2017
pmsx: 0.4673131003419756
feature_pmax foreign_suspects_2014
pmsx: 0.519895157438091
feature_pmax total_suspects_2014
pmsx: 0.4116907700877702
feature_pmax germans_2017
pmsx: 0.38064458104092536
feature_pmax graduates_without_secondary_2017
pmsx: 0.3861572790524159
feature_pmax age_to_18_2017_2012
pmsx: 0.46152665683809024
feature_pmax dwellings_2017_2012
pmsx: 0.43650639560635274
feature_pmax foreign_suspects_2013
pmsx: 0.3584931891094483
feature_pmax trade_tax_per_inh_2012
pmsx: 0.6419766984400108
feature_pmax gdp_2016
pmsx: 0.3610552871237611
feature_pmax graduates_voc_2017
pmsx: 0.38646444911569056
feature_pmax total_suspects_2017
pmsx: 0.3460927918986908
feature_pmax Absolventen/Abgänger allgemeinbildender Schulen 2017 - insgesamt ohne Externe (je 100
0 Einwohner)

pmsx: 0.32595615876885786
feature_pmax empl_service_2018
pmsx: 0.3076091118046995
feature_pmax age_75_more_2017
pmsx: 0.35783118817438975
feature_pmax ag4_60_74_2017_2012
pmsx: 0.6083802506967217
feature_pmax age_75_more_2017_2012
pmsx: 0.32027668018626576
feature_pmax age_18_24_2017
pmsx: 0.32547877993967367
feature_pmax child_day_care_2018
pmsx: 0.2643490244163826
feature_pmax population_2017
pmsx: 0.30034009431974984
feature_pmax male_2012
pmsx: 0.32031601855089153
feature_pmax empl_oth_service_2018_2012
pmsx: 0.24852927609120157
feature_pmax insolvencies_2017_2012
pmsx: 0.20365757132516063
feature_pmax protection_total_2017
pmsx: 0.20350543200290558
feature_pmax hartz_no_empl_2013
pmsx: 0.1605249840378828
feature_pmax Nr
pmsx: 0.1612982170878141
feature_pmax area_2017
pmsx: 0.2068166493381489
feature_pmax total_suspects_2012
pmsx: 0.31050631310597515
feature_pmax total_suspects_2016
pmsx: 0.16187379983341341
feature_pmax hartz_total_2018_2013
pmsx: 0.17108247200437782
feature_pmax graduates_secondary_2017
pmsx: 0.14877894554023494
feature_pmax f_crime_2018
pmsx: 0.15219606587059153
feature_pmax debt_2013
pmsx: 0.21607805282614473
feature_pmax turnout19_14
pmsx: 0.16132379008501704
feature_pmax f_crime_2012
pmsx: 0.28254481441675827
feature_pmax foreign_suspects_2017
pmsx: 0.25244928781323106
feature_pmax f_crime_2017
pmsx: 0.09818221600324856
feature_pmax birth_balance_2012
pmsx: 0.09558090120971266
feature_pmax age_to_18_2017
pmsx: 0.08630062342599532
feature_pmax dwellings_new_2012
pmsx: 0.08298203247457205
feature_pmax dwellings_2012
pmsx: 0.13747168662214862
feature_pmax population_density_2017
pmsx: 0.08852110328133143
feature_pmax graduates_without_secondary_2017_2012
pmsx: 0.14146611114200128
feature_pmax graduates_without_secondary_2012
pmsx: 0.20589961777903096
feature_pmax graduates_lower_secondary_2012
pmsx: 0.37783060194132645
feature_pmax graduates_uni_2012
pmsx: 0.33298062177223486
feature_pmax graduates_secondary_2012
pmsx: 0.07614079786741134
feature_pmax empl_service_2018_2012
pmsx: 0.06689255705138185
feature_pmax net_migration_2017
pmsx: 0.051830349310663014
feature_pmax empl_com_hotel_2018_2012
pmsx: 0.05027895551829242
feature_pmax empl_manuf_2018
pmsx: 0.058006467980172796

```
pmsx: 0.05800040790017290
feature_pmax empl_agr_2018_2012
pmsx: 0.06787483256175696
feature_pmax vehicles_2018_2013
pmsx: 0.12956352777424276
feature_pmax age_18_24_2017_2012
pmsx: 0.059701709898329826
feature_pmax unempl_male_2019_2013
pmsx: 0.02702105685236611
feature_pmax unempl_female_2019_2013
['subregion', 'turnout14', 'ove18_13', 'age_25_34_2017', 'age_35_59_2017', 'dwellings_2017',
'space_per_app_2017', 'space_per_inh_2017', 'vehicles_2018', 'graduates_lower_secondary_2017', 'gr
aduates_higher_2017', 'business_reg_2017', 'insolvencies_2017', 'empl_agr_2018',
'empl_com_hotel_2018', 'hartz_total_2018', 'hartz_foreign_2018', 'unempl_total_2019',
'unempl_male_2019', 'unempl_female_2019', 'unempl_55_64_2019', 'birth_balance_2017_2012',
'graduates_lower_secondary_2017_2012', 'graduates_secondary_2017_2012', 'dwellings_new_2017_2012',
'unempl_total_2019_2013', 'unempl_female_2019_2013', 'population_density_2012', 'age_to_18_2012',
'age_18_24_2012', 'age_35_59_2012', 'ag4_60_74_2012', 'age_75_more_2012', 'vehicles_2013',
'mining_manuf_2012', 'empl_manuf_2012', 'empl_com_hotel_2012', 'empl_service_2012',
'empl_oth_service_2012', 'unempl_total_2013', 'unempl_male_2013', 'foreign_suspects_2015',
'f_crime_2013', 'foreign_suspects_2012']
```

In [322]:

```
p
```

Out[322]:

```
subregion                              5.345279e-06
turnout14                              6.577139e-10
ove18_13                               3.402800e-06
age_25_34_2017                         1.372517e-05
age_35_59_2017                         7.938381e-04
dwellings_2017                         4.763489e-06
space_per_app_2017                     1.826875e-06
space_per_inh_2017                     6.895299e-08
vehicles_2018                          3.444644e-03
graduates_lower_secondary_2017         1.817984e-06
graduates_higher_2017                  4.716193e-04
business_reg_2017                      1.875808e-08
insolvencies_2017                      2.348688e-04
empl_agr_2018                          1.072381e-03
empl_com_hotel_2018                    1.318474e-03
hartz_total_2018                       3.551546e-04
hartz_foreign_2018                     9.724971e-05
unempl_total_2019                      1.148729e-02
unempl_male_2019                       2.424854e-02
unempl_female_2019                     2.574754e-03
unempl_55_64_2019                      6.821649e-04
birth_balance_2017_2012                7.183837e-03
graduates_lower_secondary_2017_2012    2.471338e-03
graduates_secondary_2017_2012          7.166972e-03
dwellings_new_2017_2012                8.130134e-03
unempl_total_2019_2013                 2.644346e-02
unempl_female_2019_2013                2.702106e-02
population_density_2012                1.607579e-02
age_to_18_2012                         5.210391e-10
age_18_24_2012                         2.229763e-14
age_35_59_2012                         1.071903e-02
ag4_60_74_2012                         1.092947e-09
age_75_more_2012                       8.612342e-06
vehicles_2013                          7.055586e-04
mining_manuf_2012                      1.396521e-02
empl_manuf_2012                        7.550964e-03
empl_com_hotel_2012                    1.071291e-03
empl_service_2012                      5.676659e-03
empl_oth_service_2012                  5.092879e-03
unempl_total_2013                      3.924841e-05
unempl_male_2013                       6.459955e-05
foreign_suspects_2015                  1.297292e-02
f_crime_2013                           4.492678e-03
foreign_suspects_2012                  1.738121e-02
dtype: float64
```

In [26]:

```
len(selected_features_BE)
```

Out[26]:

44

In [27]:

```
model2.summary()
```

Out[27]:

OLS Regression Results

Dep. Variable:	vot19_14	**R-squared:**	0.933
Model:	OLS	**Adj. R-squared:**	0.925
Method:	Least Squares	**F-statistic:**	113.5
Date:	Wed, 29 Jan 2020	**Prob (F-statistic):**	9.95e-183
Time:	12:54:55	**Log-Likelihood:**	-694.26
No. Observations:	401	**AIC:**	1479.
Df Residuals:	356	**BIC:**	1658.
Df Model:	44		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	179.2441	32.173	5.571	0.000	115.971	242.518
subregion	0.2111	0.046	4.621	0.000	0.121	0.301
turnout14	9.8914	1.558	6.350	0.000	6.828	12.955
ove18_13	0.8871	0.188	4.720	0.000	0.517	1.257
age_25_34_2017	-0.9815	0.223	-4.410	0.000	-1.419	-0.544
age_35_59_2017	-0.7039	0.208	-3.384	0.001	-1.113	-0.295
dwellings_2017	0.0649	0.014	4.646	0.000	0.037	0.092
space_per_app_2017	0.3533	0.073	4.853	0.000	0.210	0.497
space_per_inh_2017	-0.8240	0.150	-5.510	0.000	-1.118	-0.530
vehicles_2018	-0.0188	0.006	-2.945	0.003	-0.031	-0.006
graduates_lower_secondary_2017	-0.1260	0.026	-4.854	0.000	-0.177	-0.075
graduates_higher_2017	-0.0592	0.017	-3.529	0.000	-0.092	-0.026
business_reg_2017	-0.1388	0.024	-5.754	0.000	-0.186	-0.091
insolvencies_2017	0.4166	0.112	3.716	0.000	0.196	0.637
empl_agr_2018	-0.9002	0.273	-3.298	0.001	-1.437	-0.363
empl_com_hotel_2018	0.1774	0.055	3.238	0.001	0.070	0.285
hartz_total_2018	0.0391	0.011	3.606	0.000	0.018	0.060
hartz_foreign_2018	-0.0671	0.017	-3.942	0.000	-0.101	-0.034
unempl_total_2019	5.5426	2.182	2.541	0.011	1.252	9.833
unempl_male_2019	-2.6297	1.162	-2.263	0.024	-4.915	-0.344
unempl_female_2019	-3.1808	1.048	-3.036	0.003	-5.241	-1.120
unempl_55_64_2019	0.5143	0.150	3.427	0.001	0.219	0.809
birth_balance_2017_2012	0.0006	0.000	2.704	0.007	0.000	0.001
graduates_lower_secondary_2017_2012	0.0099	0.003	3.049	0.002	0.004	0.016
graduates_secondary_2017_2012	-0.0127	0.005	-2.705	0.007	-0.022	-0.003
dwellings_new_2017_2012	0.0013	0.000	2.662	0.008	0.000	0.002
unempl_total_2019_2013	-0.0100	0.005	-2.229	0.026	-0.019	-0.001
unempl_female_2019_2013	0.0235	0.011	2.220	0.027	0.003	0.044
population_density_2012	-0.0007	0.000	-2.419	0.016	-0.001	-0.000

age_to_18_2012	-1.4904	0.233	-6.390	0.000	-1.949	-1.032
age_18_24_2012	-2.0951	0.263	-7.966	0.000	-2.612	-1.578
age_35_59_2012	-0.5868	0.229	-2.565	0.011	-1.037	-0.137
ag4_60_74_2012	-1.5377	0.246	-6.262	0.000	-2.021	-1.055
age_75_more_2012	-1.0773	0.239	-4.515	0.000	-1.547	-0.608
vehicles_2013	0.0243	0.007	3.417	0.001	0.010	0.038
mining_manuf_2012	1.1095	0.449	2.470	0.014	0.226	1.993
empl_manuf_2012	-0.6631	0.247	-2.687	0.008	-1.148	-0.178
empl_com_hotel_2012	-0.8386	0.254	-3.298	0.001	-1.339	-0.339
empl_service_2012	-0.6869	0.247	-2.783	0.006	-1.172	-0.201
empl_oth_service_2012	-0.7031	0.249	-2.819	0.005	-1.194	-0.213
unempl_total_2013	-1.4867	0.357	-4.164	0.000	-2.189	-0.785
unempl_male_2013	1.3974	0.346	4.043	0.000	0.718	2.077
foreign_suspects_2015	0.0046	0.002	2.497	0.013	0.001	0.008
f_crime_2013	-0.0336	0.012	-2.860	0.004	-0.057	-0.010
foreign_suspects_2012	-0.0053	0.002	-2.390	0.017	-0.010	-0.001

Omnibus:	7.274	**Durbin-Watson:**	1.837
Prob(Omnibus):	0.026	**Jarque-Bera (JB):**	9.411
Skew:	0.160	**Prob(JB):**	0.00905
Kurtosis:	3.679	**Cond. No.**	5.75e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.75e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

In [30]:

```
df_4 = df4[selected_features_BE]
```

In [31]:

```
X = df_4
Y = df4["vot19_14"]
x1 = sm.add_constant(X) #Adding the constant
model_2 = sm.OLS(Y,x1).fit() # fitting the model
print(model_2.summary()) # model summary
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                vot19_14   R-squared:                       0.933
Model:                            OLS   Adj. R-squared:                  0.925
Method:                 Least Squares   F-statistic:                     113.5
Date:                Wed, 29 Jan 2020   Prob (F-statistic):          9.95e-183
Time:                        12:57:07   Log-Likelihood:                -694.26
No. Observations:                 401   AIC:                             1479.
Df Residuals:                     356   BIC:                             1658.
Df Model:                          44
Covariance Type:            nonrobust
================================================================================================

                                   coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------------------------
const                          179.2441     32.173      5.571      0.000     115.971      242
18
subregion                        0.2111      0.046      4.621      0.000       0.121        0
```

subregion	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
01						
turnout14	9.8914	1.558	6.350	0.000	6.828	12
55						
ove18_13	0.8871	0.188	4.720	0.000	0.517	1
57						
age_25_34_2017	-0.9815	0.223	-4.410	0.000	-1.419	-0
44						
age_35_59_2017	-0.7039	0.208	-3.384	0.001	-1.113	-0
95						
dwellings_2017	0.0649	0.014	4.646	0.000	0.037	0
92						
space_per_app_2017	0.3533	0.073	4.853	0.000	0.210	0
97						
space_per_inh_2017	-0.8240	0.150	-5.510	0.000	-1.118	-0
530						
vehicles_2018	-0.0188	0.006	-2.945	0.003	-0.031	-0
06						
graduates_lower_secondary_2017	-0.1260	0.026	-4.854	0.000	-0.177	-0
075						
graduates_higher_2017	-0.0592	0.017	-3.529	0.000	-0.092	-0
026						
business_reg_2017	-0.1388	0.024	-5.754	0.000	-0.186	-0
091						
insolvencies_2017	0.4166	0.112	3.716	0.000	0.196	0
37						
empl_agr_2018	-0.9002	0.273	-3.298	0.001	-1.437	-0
63						
empl_com_hotel_2018	0.1774	0.055	3.238	0.001	0.070	0
85						
hartz_total_2018	0.0391	0.011	3.606	0.000	0.018	0
60						
hartz_foreign_2018	-0.0671	0.017	-3.942	0.000	-0.101	-0
034						
unempl_total_2019	5.5426	2.182	2.541	0.011	1.252	9
33						
unempl_male_2019	-2.6297	1.162	-2.263	0.024	-4.915	-
0.344						
unempl_female_2019	-3.1808	1.048	-3.036	0.003	-5.241	-1
120						
unempl_55_64_2019	0.5143	0.150	3.427	0.001	0.219	0
09						
birth_balance_2017_2012	0.0006	0.000	2.704	0.007	0.000	
0.001						
graduates_lower_secondary_2017_2012	0.0099	0.003	3.049	0.002	0.004	0
016						
graduates_secondary_2017_2012	-0.0127	0.005	-2.705	0.007	-0.022	-0
003						
dwellings_new_2017_2012	0.0013	0.000	2.662	0.008	0.000	
0.002						
unempl_total_2019_2013	-0.0100	0.005	-2.229	0.026	-0.019	-0
001						
unempl_female_2019_2013	0.0235	0.011	2.220	0.027	0.003	
0.044						
population_density_2012	-0.0007	0.000	-2.419	0.016	-0.001	-0
000						
age_to_18_2012	-1.4904	0.233	-6.390	0.000	-1.949	-1
32						
age_18_24_2012	-2.0951	0.263	-7.966	0.000	-2.612	-1
78						
age_35_59_2012	-0.5868	0.229	-2.565	0.011	-1.037	-0
37						
ag4_60_74_2012	-1.5377	0.246	-6.262	0.000	-2.021	-1
55						
age_75_more_2012	-1.0773	0.239	-4.515	0.000	-1.547	-
0.608						
vehicles_2013	0.0243	0.007	3.417	0.001	0.010	0
38						
mining_manuf_2012	1.1095	0.449	2.470	0.014	0.226	1
93						
empl_manuf_2012	-0.6631	0.247	-2.687	0.008	-1.148	-
0.178						
empl_com_hotel_2012	-0.8386	0.254	-3.298	0.001	-1.339	-0
339						
empl_service_2012	-0.6869	0.247	-2.783	0.006	-1.172	-0
201						
empl_oth_service_2012	-0.7031	0.249	-2.819	0.005	-1.194	-0
213						

```
unempl_total_2013                -1.4867      0.357     -4.164      0.000      -2.189      -0
785
unempl_male_2013                  1.3974      0.346      4.043      0.000       0.718       2
77
foreign_suspects_2015             0.0046      0.002      2.497      0.013       0.001
0.008
f_crime_2013                     -0.0336      0.012     -2.860      0.004      -0.057      -0
10
foreign_suspects_2012            -0.0053      0.002     -2.390      0.017      -0.010      -0
001
==============================================================================
Omnibus:                        7.274   Durbin-Watson:                   1.837
Prob(Omnibus):                  0.026   Jarque-Bera (JB):                9.411
Skew:                           0.160   Prob(JB):                      0.00905
Kurtosis:                       3.679   Cond. No.                     5.75e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.75e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [335]:

```
vif = pd.DataFrame()
vif["VIF Factor"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif["features"] = X.columns
```

In [336]:

```

```

Out[336]:

	VIF Factor	features
0	30.517343	subregion
1	100.909104	turnout14
2	2.300262	ove18_13
3	1020.758799	age_25_34_2017
4	9741.947778	age_35_59_2017
5	9418.804444	dwellings_2017
6	9368.383514	space_per_app_2017
7	9805.965533	space_per_inh_2017
8	4107.226336	vehicles_2018
9	38.425765	graduates_lower_secondary_2017
10	61.133338	graduates_higher_2017
11	185.305227	business_reg_2017
12	128.418779	insolvencies_2017
13	22.537287	empl_agr_2018
14	296.605677	empl_com_hotel_2018
15	115.627115	hartz_total_2018
16	70.307820	hartz_foreign_2018
17	27928.806850	unempl_total_2019
18	9067.327337	unempl_male_2019
19	5530.198184	unempl_female_2019
20	159.547288	unempl_55_64_2019
21	1.126343	birth_balance_2017_2012
22	1.923416	graduates_lower_secondary_2017_2012
23	2.044678	graduates_secondary_2017_2012
24	1.521563	dwellings_new_2017_2012

	VIF Factor	features
25	4.419806	unempl_total_2019_2013
26	91.737292	unempl_female_2019_2013
27	11.221500	population_density_2012
28	1897.312674	age_to_18_2012
29	640.097708	age_18_24_2012
30	10120.491308	age_35_59_2012
31	2433.709662	ag4_60_74_2012
32	868.497070	age_75_more_2012
33	4465.526498	vehicles_2013
34	16.945841	mining_manuf_2012
35	7772.734453	empl_manuf_2012
36	3459.873411	empl_com_hotel_2012
37	4434.738684	empl_service_2012
38	1778.992029	empl_oth_service_2012
39	1160.559252	unempl_total_2013
40	1061.477106	unempl_male_2013
41	75.981488	foreign_suspects_2015
42	21.267132	f_crime_2013
43	76.402923	foreign_suspects_2012

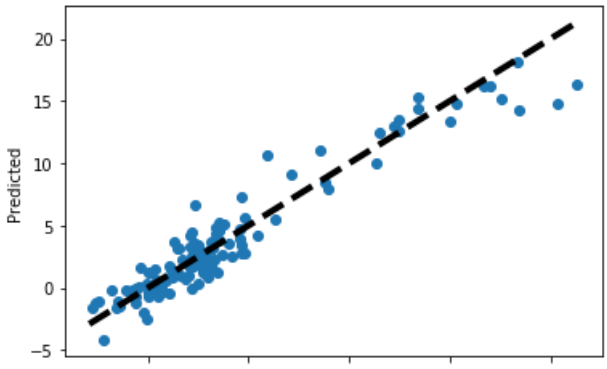
In [32]:

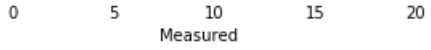
```
# model2 : with p values < 0.05 Backward elimination with  subregion and Nr

X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.3, random_state = 12)
reg = LinearRegression()
m_2 = reg.fit(X_train, Y_train)
#Predict the model on train
pred_train = m_2.predict(X_train)
print("MSE on train: ", mean_squared_error(Y_train,pred_train))
print("RSquared: ",m_2.score(X_train,Y_train))
#Predict on test
pred_test = m_2.predict(X_test)
print("MSE on test: ", mean_squared_error(Y_test,pred_test))
print("RSquared: ",m_2.score(X_test,Y_test))
```

```
MSE on train:  1.8526466142349656
RSquared:  0.935216740673382
MSE on test:  2.3068059689022884
RSquared:  0.9124425045324986
```

In [271]:

```
fig, ax = plt.subplots()
ax.scatter(Y_test, pred_test)
ax.plot([Y.min(), Y.max()], [Y.min(), Y.max()], 'k--', lw=4)
ax.set_xlabel('Measured')
ax.set_ylabel('Predicted')
plt.show()
```





In [38]:

```
values = pd.DataFrame(pred_test,Y_test,)
values
```

Out[38]:

	0
vot19_14	
2.400414	1.944227
18.334998	18.200895
-0.724483	-0.171693
0.060316	0.004437
-0.329070	1.663082
...	...
2.524834	3.417599
3.027529	1.544795
2.114120	1.584297
2.299495	2.484344
0.252615	0.247788

121 rows × 1 columns

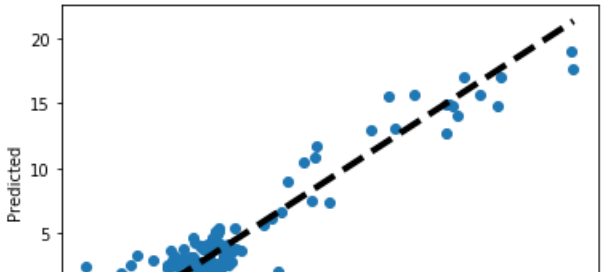
In [272]:

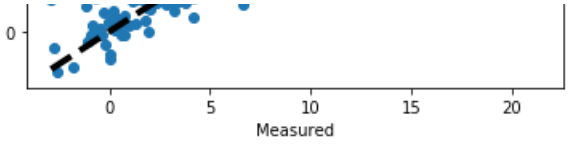
```
# model2 with backward elimination with out subregion and Nr
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.3, random_state = 2)
reg = LinearRegression()
m2 = reg.fit(X_train, Y_train)
#Predict the model on train
pred_train = m2.predict(X_train)
print("MSE on train: ", mean_squared_error(Y_train,pred_train))
print("RSquared: ",m2.score(X_train,Y_train))
#Predict on test
pred_test = m2.predict(X_test)
print("MSE on test: ", mean_squared_error(Y_test,pred_test))
print("RSquared: ",m2.score(X_test,Y_test))
```

```
MSE on train:  1.7747006616235133
RSquared:  0.9386584521015146
MSE on test:  2.4192473637960967
RSquared:  0.9062132804166938
```

In [273]:

```
fig, ax = plt.subplots()
ax.scatter(Y_test, pred_test)
ax.plot([Y.min(), Y.max()], [Y.min(), Y.max()], 'k--', lw=4)
ax.set_xlabel('Measured')
ax.set_ylabel('Predicted')
plt.show()
```





model3: Using recursive feature extraction

In [207]:

```
df5 = df1.copy()
```

In [208]:

```
X= df5.drop("vot19_14", axis = 1)
y = df5["vot19_14"]
model3 = LinearRegression()
#Initializing RFE model
rfe = RFE(model3, 150)
#Transforming data using RFE
X_rfe = rfe.fit_transform(X,y)
#Fitting the data to model
model3.fit(X_rfe,y)
print(rfe.support_)
print(rfe.ranking_)
```

```
[ True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True]
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
```

In [209]:

```
#no of features
nof_list=np.arange(1,150)
high_score=0
#Variable to store the optimum features
nof=0
score_list =[]
for n in range(len(nof_list)):
    X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.3, random_state = 0)
    model3 = LinearRegression()
    rfe = RFE(model,nof_list[n])
    X_train_rfe = rfe.fit_transform(X_train,y_train)
    X_test_rfe = rfe.transform(X_test)
    model3.fit(X_train_rfe,y_train)
    score = model3.score(X_test_rfe,y_test)
    score_list.append(score)
    if(score>high_score):
        high_score = score
        nof = nof_list[n]
print("Optimum number of features: %d" %nof)
print("Score with %d features: %f" % (nof, high_score))
```

```
Optimum number of features: 50
Score with 50 features: 0.882529
```

In [210]:

```
cols = list(X.columns)
model3 = LinearRegression()
#Initializing RFE model
rfe = RFE(model3, 50)
#Transforming data using RFE
X_rfe = rfe.fit_transform(X,y)
#Fitting the data to model
model3.fit(X_rfe,y)
temp = pd.Series(rfe.support_,index = cols)
selected_features_rfe = temp[temp==True].index
print(selected_features_rfe)
```

```
Index(['subregion', 'turnout14', 'turnout19_14', 'debt_2013', 'debt_2014',
       'debt_2017', 'debt_2018', 'ove18_13', 'foreigners_2017',
       'birth_balance_2017', 'age_to_18_2017', 'age_18_24_2017',
       'age_25_34_2017', 'age_35_59_2017', 'ag4_60_74_2017',
       'age_75_more_2017', 'protection_open_2017', 'protection_accepted_2017',
       'protection_rejected_2017', 'dwellings_new_2017',
       'graduates_without_secondary_2017', 'graduates_lower_secondary_2017',
       'graduates_secondary_2017', 'graduates_higher_2017',
       'business_reg_2017', 'insolvencies_2017', 'empl_agr_2018',
       'empl_com_hotel_2018', 'unempl_total_2019', 'unempl_male_2019',
       'unempl_female_2019', 'unempl_55_64_2019', 'birth_balance_2012',
       'age_to_18_2012', 'age_18_24_2012', 'age_35_59_2012', 'ag4_60_74_2012',
       'age_75_more_2012', 'graduates_without_secondary_2012',
       'graduates_lower_secondary_2012', 'graduates_secondary_2012',
       'graduates_uni_2012', 'mining_manuf_2012', 'insolvencies_2012',
       'empl_manuf_2012', 'empl_com_hotel_2012', 'empl_service_2012',
       'empl_oth_service_2012', 'unempl_total_2013', 'unempl_male_2013'],
      dtype='object')
```

In [211]:

```
df_5 = df5[selected_features_rfe]
```

In [213]:

```

```

Out[213]:

	subregion	turnout14	turnout19_14	debt_2013	debt_2014	debt_2017	debt_2018	ove18_13	foreigners_2017	birth_balance_2017
0	1.0	0.357400	0.205520	16.41	16.40	16.21	16.24	-0.17	13.3	-1.7
1	1.0	0.402589	0.186015	12.04	12.03	12.16	11.96	-0.08	11.3	0.1
2	1.0	0.376398	0.169726	15.25	15.59	15.04	14.76	-0.49	9.7	-3.6
3	1.0	0.453649	0.028556	16.61	16.94	17.80	18.09	1.48	11.0	-5.4
4	1.0	0.397193	0.146915	12.52	12.80	12.78	12.84	0.32	5.6	-5.3
5	1.0	0.463842	0.138119	10.16	10.29	10.17	10.23	0.07	7.5	-3.0
6	1.0	0.411905	0.176788	10.09	10.32	9.82	9.93	-0.16	6.8	-3.6
7	1.0	0.424788	0.159476	10.63	10.80	10.62	10.47	-0.16	5.5	-6.9
8	1.0	0.457528	0.172180	9.67	9.64	9.60	9.62	-0.05	10.1	-2.2
9	1.0	0.469130	0.167735	8.67	8.88	8.95	8.94	0.27	4.7	-4.0
10	1.0	0.459591	0.165319	9.16	9.17	9.41	9.38	0.22	5.0	-3.6
11	1.0	0.421130	0.177120	10.45	10.67	10.64	10.61	0.16	5.1	-3.8
12	1.0	0.423453	0.171872	10.11	10.08	9.86	9.82	-0.29	8.1	-2.2
13	1.0	0.420752	0.156371	11.43	11.66	12.18	12.15	0.72	6.6	-3.7
14	1.0	0.503786	0.150481	7.99	7.93	7.63	7.55	-0.44	7.0	-3.1
15	2.0	0.435025	0.183690	10.92	10.81	10.61	10.62	-0.30	16.2	1.9
16	3.0	0.512913	0.128824	10.62	10.47	9.61	9.67	-0.95	9.7	-2.1
17	3.0	0.459825	0.051813	12.16	12.51	13.24	13.49	1.33	16.9	-3.3
18	3.0	0.412644	0.152323	8.17	7.99	7.69	7.74	-0.43	14.3	-1.7

	subregion	turnout14	turnout19_14	debt_2013	debt_2014	debt_2017	debt_2018	ove18_13	foreigners_2017	birth_balance_2017
19	3.0	0.477766	0.129914	9.02	8.98	8.85	8.74	-0.28	6.2	-0.8
20	3.0	0.383769	0.200491	12.74	12.88	12.47	12.51	-0.23	8.9	-8.5
21	3.0	0.423608	0.160607	10.97	11.19	11.04	11.21	0.24	6.6	-5.5
22	3.0	0.481927	0.079024	10.80	10.98	10.66	10.68	-0.12	5.9	-6.7
23	3.0	0.468641	0.150865	9.78	10.00	10.33	10.41	0.63	7.2	-2.9
24	3.0	0.525676	0.118711	9.13	9.22	9.10	9.09	-0.04	5.9	-4.4
25	3.0	0.502767	0.099959	9.82	9.73	9.58	9.57	-0.25	8.1	-3.6
26	3.0	0.486508	0.153106	12.01	11.93	11.50	11.45	-0.56	13.1	-1.2
27	3.0	0.495989	0.125992	9.23	9.28	9.41	9.39	0.16	7.8	-2.9
28	3.0	0.491912	0.092370	12.23	12.40	12.78	12.93	0.70	9.7	-5.6
29	3.0	0.484514	0.139769	10.98	11.07	10.80	10.65	-0.33	7.6	-5.7
...	...	...	...	...	...	...	...	...	...	...
371	15.0	0.407577	0.127519	11.99	12.42	12.39	12.34	0.35	3.0	-8.4
372	15.0	0.479496	0.068303	11.74	11.82	11.88	11.94	0.20	3.4	-6.6
373	15.0	0.452224	0.084795	11.27	11.62	12.22	12.16	0.89	2.9	-9.7
374	15.0	0.447176	0.123932	11.30	11.45	11.60	11.59	0.29	3.5	-5.9
375	15.0	0.405973	0.094412	12.33	12.54	12.95	13.10	0.77	3.1	-9.6
376	15.0	0.427994	0.115690	11.92	12.15	12.42	12.33	0.41	3.7	-6.6
377	15.0	0.467126	0.108452	9.96	10.29	10.48	10.56	0.60	3.2	-7.5
378	16.0	0.479950	0.127485	11.48	11.45	11.17	11.19	-0.29	7.4	-1.4
379	16.0	0.435021	0.132615	11.62	11.90	11.77	11.85	0.23	5.8	-6.6
380	16.0	0.522503	0.127272	5.81	5.70	5.94	5.91	0.10	8.9	0.9
381	16.0	0.451281	0.113586	10.03	10.11	10.55	10.81	0.78	7.0	-9.3
382	16.0	0.499509	0.130687	10.79	10.49	10.75	10.74	-0.05	8.2	-2.7
383	16.0	0.516287	0.079205	11.87	12.16	12.06	11.91	0.04	7.9	-5.5
384	16.0	0.573239	0.085326	6.19	6.27	6.32	6.14	-0.05	3.3	-2.5
385	16.0	0.485793	0.102730	9.47	9.82	10.19	10.42	0.95	4.2	-6.4
386	16.0	0.544328	0.081478	7.53	7.68	8.05	8.01	0.48	2.8	-5.4
387	16.0	0.496123	0.098784	9.80	9.95	9.91	10.21	0.41	3.5	-5.7
388	16.0	0.516722	0.077842	10.88	10.97	11.40	11.56	0.68	2.8	-8.6
389	16.0	0.521090	0.090600	8.92	9.11	9.54	9.54	0.62	3.2	-6.3
390	16.0	0.519206	0.079432	9.66	9.73	10.45	10.59	0.93	5.4	-5.5
391	16.0	0.537037	0.074817	9.03	9.21	9.47	9.54	0.51	2.6	-4.5
392	16.0	0.524583	0.121694	8.01	8.25	8.35	8.59	0.58	2.9	-6.5
393	16.0	0.528120	0.106144	9.18	9.21	9.49	9.57	0.39	5.3	-5.2
394	16.0	0.565730	0.092031	9.30	9.36	9.49	9.46	0.16	3.5	-4.1
395	16.0	0.473913	0.100873	8.83	8.81	8.87	8.93	0.10	4.1	-8.5
396	16.0	0.508175	0.101750	7.45	7.81	8.18	8.23	0.78	2.7	-8.5
397	16.0	0.584836	0.075714	7.40	7.40	7.43	7.50	0.10	2.9	-4.5
398	16.0	0.541176	0.093087	7.48	7.56	7.52	7.47	-0.01	3.3	-6.7
399	16.0	0.553222	0.088169	7.42	7.77	7.80	7.76	0.34	2.3	-7.7
400	16.0	0.483017	0.082643	8.29	8.66	8.87	8.95	0.66	3.2	-7.9

401 rows × 50 columns

In [212]:

```
X = df_5
Y = df5["vot19_14"]
x1 = sm.add_constant(X) #Adding the constant
model3 = sm.OLS(Y,x1).fit() # fitting the model
print(model3.summary()) # model summary
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                vot19_14   R-squared:                       0.924
Model:                            OLS   Adj. R-squared:                  0.913
Method:                 Least Squares   F-statistic:                     86.65
Date:                Mon, 27 Jan 2020   Prob (F-statistic):          2.85e-167
Time:                        21:31:40   Log-Likelihood:                -721.79
No. Observations:                 401   AIC:                             1544.
Df Residuals:                     351   BIC:                             1743.
Df Model:                          49
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975
------------------------------------------------------------------------------------------------
const                           26.3911    137.642      0.192      0.848    -244.316     297.09
subregion                        0.2039      0.053      3.865      0.000       0.100       0.30
turnout14                        4.5891      2.901      1.582      0.115      -1.117      10.29
turnout19_14                    -4.3366      4.616     -0.939      0.348     -13.415       4.74
debt_2013                       -0.5717      0.501     -1.142      0.254      -1.556       0.41
debt_2014                        1.0318      0.815      1.267      0.206      -0.570       2.63
debt_2017                       -0.7914      0.795     -0.995      0.320      -2.356       0.77
debt_2018                        0.2807      0.508      0.553      0.581      -0.718       1.27
ove18_13                         0.8524      0.335      2.548      0.011       0.194       1.51
foreigners_2017                 -0.1232      0.042     -2.960      0.003      -0.205      -0.04
birth_balance_2017              -0.2002      0.127     -1.578      0.115      -0.450       0.04
age_to_18_2017                   0.7102      1.163      0.610      0.542      -1.578       2.99
age_18_24_2017                  -1.3520      1.146     -1.179      0.239      -3.607       0.90
age_25_34_2017                  -1.2289      1.145     -1.073      0.284      -3.481       1.02
age_35_59_2017                  -1.4633      1.132     -1.293      0.197      -3.690       0.76
ag4_60_74_2017                  -0.6139      1.115     -0.550      0.582      -2.808       1.58
age_75_more_2017                -0.6423      1.118     -0.574      0.566      -2.841       1.55
protection_open_2017             1.9861      0.750      2.650      0.008       0.512       3.46
protection_accepted_2017         1.9684      0.748      2.630      0.009       0.497       3.44
protection_rejected_2017         1.9895      0.748      2.660      0.008       0.519       3.46
dwellings_new_2017               0.0978      0.067      1.452      0.147      -0.035       0.23
graduates_without_secondary_2017 -0.5100     1.079     -0.473      0.637      -2.633       1.61
graduates_lower_secondary_2017   -0.6096     1.076     -0.567      0.571      -2.726       1.50
graduates_secondary_2017         -0.5832     1.073     -0.543      0.587      -2.694       1.52
graduates_higher_2017            -0.5820     1.073     -0.542      0.588      -2.692       1.52
business_reg_2017                -0.1603     0.025     -6.318      0.000      -0.210      -0.11
insolvencies_2017                 0.5181     0.120      4.319      0.000       0.282       0.75
empl_agr_2018                    -0.5051     0.281     -1.796      0.073      -1.058       0.04
empl_com_hotel_2018               0.1436     0.059      2.422      0.016       0.027       0.26
unempl_total_2019                 2.8269     2.367      1.195      0.233      -1.828       7.48
unempl_male_2019                 -1.4261     1.263     -1.129      0.260      -3.910       1.05
unempl_female_2019               -1.5164     1.138     -1.333      0.184      -3.754       0.72
unempl_55_64_2019                 0.4609     0.158      2.923      0.004       0.151       0.77
birth_balance_2012               -0.3188     0.154     -2.070      0.039      -0.622      -0.01
age_to_18_2012                   -2.1907     0.316     -6.929      0.000      -2.812      -1.56
age_18_24_2012                   -1.2463     0.430     -2.898      0.004      -2.092      -0.40
age_35_59_2012                   -0.5382     0.280     -1.920      0.056      -1.089       0.01
ag4_60_74_2012                   -1.2698     0.290     -4.382      0.000      -1.840      -0.70
age_75_more_2012                 -1.4948     0.365     -4.093      0.000      -2.213      -0.77
graduates_without_secondary_2012  0.5246     1.022      0.513      0.608      -1.485       2.53
graduates_lower_secondary_2012    0.5494     1.024      0.537      0.592      -1.464       2.56
graduates_secondary_2012          0.6049     1.021      0.593      0.554      -1.403       2.61
graduates_uni_2012                0.5730     1.021      0.561      0.575      -1.435       2.58
mining_manuf_2012                 1.3775     0.516      2.670      0.008       0.363       2.39
insolvencies_2012                 0.1412     0.222      0.637      0.525      -0.295       0.57
empl_manuf_2012                  -0.3827     0.260     -1.473      0.142      -0.893       0.12
empl_com_hotel_2012              -0.5137     0.268     -1.914      0.056      -1.042       0.01
empl_service_2012                -0.4151     0.260     -1.596      0.111      -0.927       0.09
empl_oth_service_2012            -0.3649     0.261     -1.400      0.162      -0.877       0.14
unempl_total_2013                -1.0337     0.401     -2.581      0.010      -1.821      -0.24
unempl_male_2013                  1.1860     0.392      3.029      0.003       0.416       1.95
==============================================================================
Omnibus:                        6.246   Durbin-Watson:                   1.819
Prob(Omnibus):                  0.044   Jarque-Bera (JB):                6.522
Skew:                           0.221   Prob(JB):                       0.0384
Kurtosis:                       3.441   Cond. No.                     1.00e+16
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 8.61e-26. This might indicate that there are
```

[2] The smallest eigenvalue is 8.61e-26. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

```
C:\Users\SOURAV CH\Anaconda3\lib\site-packages\numpy\core\fromnumeric.py:2389: FutureWarning: Meth
od .ptp is deprecated and will be removed in a future version. Use numpy.ptp instead.
  return ptp(axis=axis, out=out, **kwargs)
```

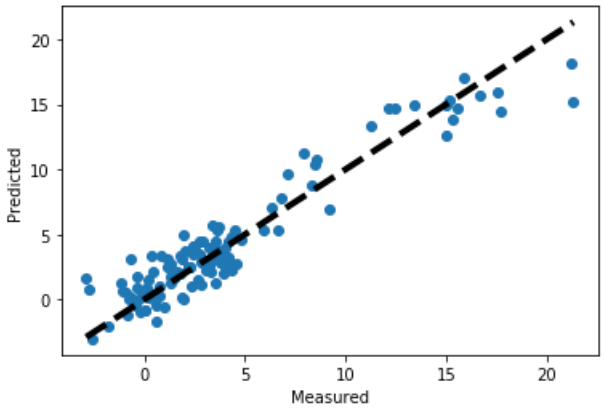
In [240]:

```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.3, random_state = 2)
reg = LinearRegression()
m3 = reg.fit(X_train, Y_train)
#Predict the model on train
pred_train = m3.predict(X_train)
print("MSE on train: ", mean_squared_error(Y_train,pred_train))
print("RSquared: ",m3.score(X_train,Y_train))
#Predict on test
pred_test = m3.predict(X_test)
print("MSE on test: ", mean_squared_error(Y_test,pred_test))
print("RSquared: ",m3.score(X_test,Y_test))
```

```
MSE on train:  2.121532024377327
RSquared:  0.9266704176621758
MSE on test:  2.6026600970093887
RSquared:  0.8991029374086554
```

In [249]:

```
fig, ax = plt.subplots()
ax.scatter(Y_test, pred_test)
ax.plot([Y.min(), Y.max()], [Y.min(), Y.max()], 'k--', lw=4)
ax.set_xlabel('Measured')
ax.set_ylabel('Predicted')
plt.show()
```



In [248]:

```
output = pd.DataFrame(pred_test,Y_test,)
output
```

Out[248]:

	0
vot19_14	
1.762276	3.421405
12.436741	14.674870
1.762954	1.946727
3.694157	5.564009
3.027529	2.302756
1.319050	1.295900
4.514345	5.390341

	0
8.472181	10.389270
vot19_14	
3.185247	2.593146
7.942587	11.313399
-0.386746	1.705652
4.207446	4.434960
17.741770	14.468807
1.895436	3.122920
-2.899875	1.650627
12.114440	14.665437
1.137629	2.529810
0.360833	3.360205
-0.728325	3.059134
3.535368	1.300015
-0.395556	0.835237
4.346890	4.786379
0.609086	-1.635894
0.551817	-0.501085
3.475091	3.145107
6.652091	5.305238
2.396875	3.800095
3.124034	2.932013
3.677925	3.466290
2.140122	2.543653
...	...
0.160638	0.234240
1.921189	4.915851
-0.481812	-0.484339
1.366896	2.145843
3.168329	2.860579
0.149390	0.110983
0.948635	-0.537281
1.952564	0.014442
2.524834	4.081914
0.751892	1.007971
4.193089	2.970592
8.565292	10.737535
2.812249	1.125089
1.288777	2.705431
8.346655	8.747109
21.251311	18.179521
0.576312	0.459109
3.421032	3.953887
6.761899	7.792657
-1.808341	-2.053493
2.391247	3.438978
3.906186	1.971982
2.883327	4.493312
-0.329070	-0.015836
3.404176	5.642970
1.590523	2.158971

4.188082	3.279752
0.179044	1.530986
0.766647	0.249986
0.028835	-0.837450

121 rows × 1 columns

Model 4 : using constant Feature elimination

In [250]:

```
df6 = df1.copy()

from sklearn.model_selection import train_test_split
from sklearn.feature_selection import VarianceThreshold
```

In [ ]:

```
df6.drop(columns = {'subregion',"Nr"}, inplace = True)
```

In [251]:

```
X = df6.drop("vot19_14", axis = 1)
Y = df6["vot19_14"]
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.3, random_state = 10)
constant_filter = VarianceThreshold(threshold=1)
constant_filter.fit(X_train)
```

Out[251]:

```
VarianceThreshold(threshold=1)
```

In [252]:

```
constant_columns = [column for column in X_train.columns
                    if column not in X_train.columns[constant_filter.get_support()]]

print(len(constant_columns))
```

```
6
```

In [253]:

```
for column in constant_columns:
    print(column)
```

```
turnout14
turnout19
turnout19_14
ove18_13
mining_manuf_2012
insolvencies_2012
```

In [254]:

```
X_train = constant_filter.transform(X_train)
X_test = constant_filter.transform(X_test)

X_train.shape, X_test.shape
```

Out[254]:

```
((280, 141), (121, 141))
```

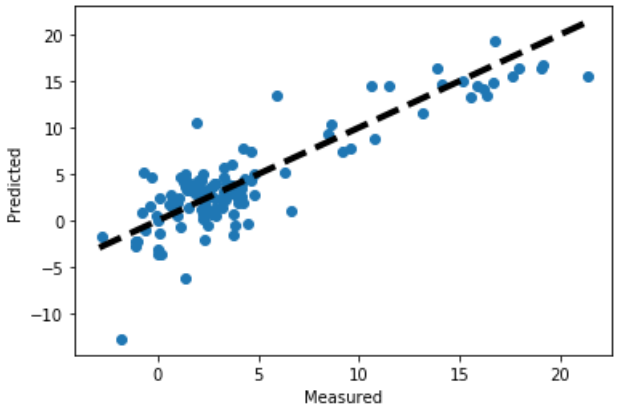
In [255]:

```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.3, random_state = 10)
reg = LinearRegression()
m4 = reg.fit(X_train, Y_train)
#Predict the model on train
pred_train = m4.predict(X_train)
print("MSE on train: ", mean_squared_error(Y_train,pred_train))
print("RSquared: ",m4.score(X_train,Y_train))
#Predict on test
pred_test = m4.predict(X_test)
print("MSE on test: ", mean_squared_error(Y_test,pred_test))
print("RSquared: ",m4.score(X_test,Y_test))
```

```
MSE on train:  1.0913312148592678
RSquared:  0.9616670989671317
MSE on test:  7.255112880255219
RSquared:  0.73243246686298
```

In [256]:

```
fig, ax = plt.subplots()
ax.scatter(Y_test, pred_test)
ax.plot([Y.min(), Y.max()], [Y.min(), Y.max()], 'k--', lw=4)
ax.set_xlabel('Measured')
ax.set_ylabel('Predicted')
plt.show()
```



Model 5 : Using highly correlated features

In [69]:

```
df7 = df1.copy()
X = df7.drop("vot19_14", axis = 1)
y = df7["vot19_14"]
```

In [276]:

```
from sklearn.feature_selection import SelectFromModel
from sklearn.linear_model import LogisticRegression
X_norm = MinMaxScaler().fit_transform(X)
embeded_lr_selector = SelectFromModel(LogisticRegression(penalty="l1"), max_features=num_feats)
embeded_lr_selector.fit(X_norm, y)

embeded_lr_support = embeded_lr_selector.get_support()
embeded_lr_feature = X.loc[:,embeded_lr_support].columns.tolist()
print(str(len(embeded_lr_feature)), 'selected features')
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-276-9f61eafa0bb0> in <module>
      3 X_norm = MinMaxScaler().fit_transform(x)
      4 embeded_lr_selector = SelectFromModel(LinearRegression(), max_features=numerical_feats)
----> 5 embeded_lr_selector.fit(X_norm, y)
```

```
      6
      7 embeded_lr_support = embeded_lr_selector.get_support()

~\Anaconda3\lib\site-packages\sklearn\feature_selection\from_model.py in fit(self, X, y,
**fit_params)
    184                 raise TypeError("'max_features' should be an integer between"
    185                                 " 0 and {} features. Got {!r} instead."
--> 186                                 .format(X.shape[1], self.max_features))
    187             elif self.max_features < 0 or self.max_features > X.shape[1]:
    188                 raise ValueError("'max_features' should be 0 and {} features."

TypeError: 'max_features' should be an integer between 0 and 147 features. Got Index(['Nr', 'subre
gion', 'vot19_14', 'turnout14', 'turnout19', 'turnout19_14',
       'debt_2013', 'debt_2014', 'debt_2015', 'debt_2016',
       ...
       'f_crime_2015', 'total_suspects_2014', 'foreign_suspects_2014',
       'f_crime_2014', 'total_suspects_2013', 'foreign_suspects_2013',
       'f_crime_2013', 'total_suspects_2012', 'foreign_suspects_2012',
       'f_crime_2012'],
      dtype='object', length=148) instead.
```

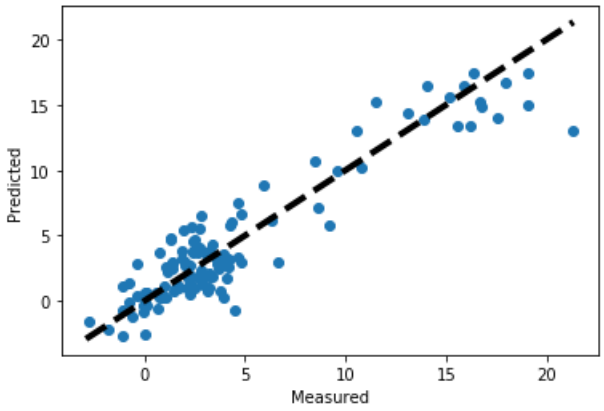
In [36]:

```
X = df2.drop("vot19_14", axis = 1)
y = df2["vot19_14"]
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.3, random_state = 10)
reg = LinearRegression()
m5 = reg.fit(X_train, y_train)
#Predict the model on train
pred_train = m5.predict(X_train)
print("MSE on train: ", mean_squared_error(y_train,pred_train))
print("RSquared: ",m5.score(X_train,y_train))
#Predict on test
pred_test = m5.predict(X_test)
print("MSE on test: ", mean_squared_error(y_test,pred_test))
print("RSquared: ",m5.score(X_test,y_test))
```

```
MSE on train:  3.4542165356646835
RSquared:  0.8786709856688132
MSE on test:  3.7889891592366176
RSquared:  0.8602626176666476
```

In [40]:

```
fig, ax = plt.subplots()
ax.scatter(y_test, pred_test)
ax.plot([y.min(), y.max()], [y.min(), y.max()], 'k--', lw=4)
ax.set_xlabel('Measured')
ax.set_ylabel('Predicted')
plt.show()
```



log Transformed model

In [57]:

```
df8 = df1.copy()
```

In [58]:

```
df8["vot19_14_log"] = np.log(df8["vot19_14"])
```

```
C:\Users\SOURAV CH\Anaconda3\lib\site-packages\pandas\core\series.py:853: RuntimeWarning: invalid
value encountered in log
  result = getattr(ufunc, method)(*inputs, **kwargs)
```

In [60]:

```
df8["vot19_14_log"].describe()
```

Out[60]:

```
count    355.000000
mean       1.091704
std        1.171142
min       -3.664176
25%        0.627194
50%        1.164356
75%        1.687447
max        3.060295
Name: vot19_14_log, dtype: float64
```

In [67]:

```
df8.fillna(df8.median(),inplace = True)
df8.drop(columns = {"vot19_14"},inplace = True)
```

In [66]:

```
df8.isnull().sum().sort_values(ascending = False)
```

Out[66]:

```
vot19_14_log             0
hartz_foreign_2018       0
hartz_total_2018         0
empl_oth_service_2018    0
empl_service_2018        0
                        ..
age_25_34_2012           0
age_18_24_2012           0
age_to_18_2012           0
net_migration_2012       0
Nr                       0
Length: 149, dtype: int64
```

In [68]:

```
X = df8.drop("vot19_14_log", axis = 1)
Y = df8["vot19_14_log"]

X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.3, random_state = 2)
reg = LinearRegression()
m6 = reg.fit(X_train, Y_train)
#Predict the model on train
pred_train = m6.predict(X_train)
print("MSE on train: ", mean_squared_error(Y_train,pred_train))
print("RSquared: ",m6.score(X_train,Y_train))
#Predict on test
pred_test = m6.predict(X_test)
print("MSE on test: ", mean_squared_error(Y_test,pred_test))
print("RSquared: ",m6.score(X_test,Y_test))
```

```
MSE on train:  0.22660811434655595
RSquared:  0.8103182575219241
MSE on test:  1.095344969974244
```

```
RSquared:  0.12071484898622765
```

In [ ]:

```
```