Checkpoint #1

Skladišta i rudarenje podataka

Dataset

- Koristi se dataset: "Election data 2019."
- Predstavlja rezultate parlamentarnih izbora u Portugalu, 2019. godine
- Prati promjene rezultata po teritorijima kroz vrijeme te ostale metrike koje će biti navedene u nastavku

Prvih 5 redaka dataseta

	TimeElapsed	time	territoryName	totalMandates	\
0	0	2019-10-06 20:10:02	Território Nacional	0	
1	0	2019-10-06 20:10:02	Território Nacional	0	
2	0	2019-10-06 20:10:02	Território Nacional	0	
3	0	2019-10-06 20:10:02	Território Nacional	0	
4	0	2019-10-06 20:10:02	Território Nacional	0	

	availableMandates	numParishes	numParishesApproved	blankVotes	\
0	226	3092	1081	9652	
1	226	3092	1081	9652	
2	226	3092	1081	9652	
3	226	3092	1081	9652	
4	226	3092	1081	9652	

	blankVotesPercentage	nullVotes	...	pre.votersPercentage	\
0	2.5	8874	...	52.66	
1	2.5	8874	...	52.66	
2	2.5	8874	...	52.66	
3	2.5	8874	...	52.66	
4	2.5	8874	...	52.66	

	pre.subscribedVoters	pre.totalVoters	Party	Mandates	Percentage	\
0	813743	428546	PS	0	38.29	
1	813743	428546	PPD/PSD	0	33.28	
2	813743	428546	B.E.	0	6.81	
3	813743	428546	CDS-PP	0	4.90	
4	813743	428546	PCP-PEV	0	4.59	

	validVotesPercentage	Votes	Hondt	FinalMandates
0	40.22	147993	94	106
1	34.95	128624	81	77
2	7.15	26307	16	19
3	5.14	18923	12	5
4	4.83	17757	11	12

[5 rows x 28 columns]

Stupci dataseta

```
# Ispis imena stupaca
print(data.columns.values)
```

```
['TimeElapsed' 'time' 'territoryName' 'totalMandates' 'availableMandates'
 'numParishes' 'numParishesApproved' 'blankVotes' 'blankVotesPercentage'
 'nullVotes' 'nullVotesPercentage' 'votersPercentage' 'subscribedVoters'
 'totalVoters' 'pre.blankVotes' 'pre.blankVotesPercentage' 'pre.nullVotes'
 'pre.nullVotesPercentage' 'pre.votersPercentage' 'pre.subscribedVoters'
 'pre.totalVoters' 'Party' 'Mandates' 'Percentage' 'validVotesPercentage'
 'Votes' 'Hondt' 'FinalMandates']
```

Dataset se sastoji od stupaca:

- Time elapsed, time, territoryName, totalMandates,
- availableMandates, numParishes,
- numParishesApproved, blankVotes,
- blankVotesPercentage, nullVotes,
- nullVotesPercentage, votersPercentage,
- subscribedVoters, totalVoters, pre.blankVotes,
- pre.blankVotesPercentage, pre.nullVotes,
- pre.nullVotesPercentage, pre.votersPercentage,
- pre.subscribedVoters, pre.totalVoters, Party,
- Mandates, Percentage, validVotesPercentage,
- Votes, Hondt, FinalMandates

Veličina dataseta

- Dataset se sastoji od 21643 retka i 28 stupaca

```
(21643, 28)
```

Null vrijednosti

- Dataset nema niti 1 null vrijednost

```
# Ispis imena stupaca i nedostajućih vrijednosti
print(data.isna().sum())
```

```
TimeElapsed                0
time                       0
territoryName              0
totalMandates              0
availableMandates          0
numParishes                0
numParishesApproved        0
blankVotes                 0
blankVotesPercentage       0
nullVotes                  0
nullVotesPercentage        0
votersPercentage           0
subscribedVoters           0
totalVoters                0
pre.blankVotes             0
pre.blankVotesPercentage   0
pre.nullVotes              0
pre.nullVotesPercentage    0
pre.votersPercentage       0
pre.subscribedVoters       0
pre.totalVoters            0
Party                      0
Mandates                   0
Percentage                 0
validVotesPercentage       0
Votes                      0
Hondt                      0
FinalMandates              0
dtype: int64
```

Jedinstvene vrijednosti

- Svaki stupac ima između 17 i 4029 unique vrijednosti

```
# Ispis broja jedinstvenih vrijednosti po stupcima
print(data.nunique())
```

```
TimeElapsed                  54
time                         54
territoryName                21
totalMandates                62
availableMandates            69
numParishes                  20
numParishesApproved         219
blankVotes                  329
blankVotesPercentage        146
nullVotes                   331
nullVotesPercentage         107
votersPercentage            282
subscribedVoters            335
totalVoters                 336
pre.blankVotes              323
pre.blankVotesPercentage    130
pre.nullVotes               329
pre.nullVotesPercentage      90
pre.votersPercentage        278
pre.subscribedVoters        331
pre.totalVoters             331
Party                        21
Mandates                     67
Percentage                 1363
validVotesPercentage       1387
Votes                      4029
Hondt                        41
FinalMandates                17
dtype: int64
```

Tipovi podataka

- Tipovi podataka u datasetu su: int64, float64 i object

```
# Ispis tipova podataka po stupcima (analiza kvanitativnih i kvalitativnih varijabli)
print(data.dtypes)
```

```
TimeElapsed                 int64
time                       object
territoryName              object
totalMandates               int64
availableMandates           int64
numParishes                 int64
numParishesApproved         int64
blankVotes                  int64
blankVotesPercentage      float64
nullVotes                   int64
nullVotesPercentage       float64
votersPercentage          float64
subscribedVoters            int64
totalVoters                 int64
pre.blankVotes              int64
pre.blankVotesPercentage  float64
pre.nullVotes               int64
pre.nullVotesPercentage   float64
pre.votersPercentage      float64
pre.subscribedVoters        int64
pre.totalVoters             int64
Party                      object
Mandates                    int64
Percentage                float64
validVotesPercentage      float64
Votes                       int64
Hondt                       int64
FinalMandates               int64
dtype: object
```

Broj jedinstvenih vrijednosti po stupcima

```
TimeElapsed
135    403
140    403
150    403
155    403
160    403
165    403
170    403
175    403
180    403
185    403
190    403
195    403
200    403
205    403
210    403
215    403
220    403
225    403
230    403
235    403
240    403
245    403
250    403
255    403
260    403
145    403
265    403
70     403
75     403
35     403
40     403
45     403
50     403
55     403
60     403
65     403
130    403
80     403
85     403
90     403
95     403
100    403
105    403
110    403
115    403
120    403
125    403
20     386
10     386
15     386
5      386
25     386
30     386
0      386
Name: count, dtype: int64
```

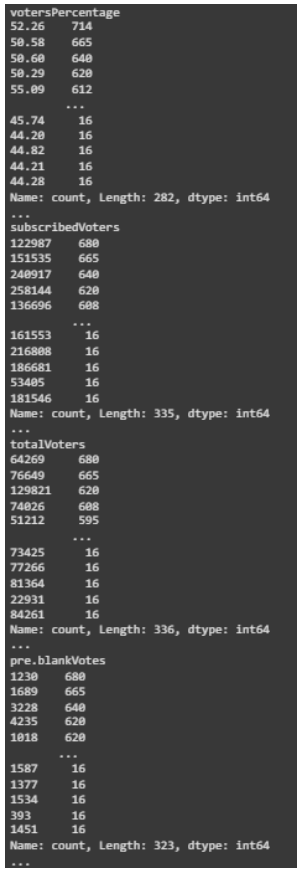
```
Name: count, dtype: int64
...
time
2019-10-06 22:25:01    403
2019-10-06 22:30:01    403
2019-10-06 22:40:02    403
2019-10-06 22:45:01    403
2019-10-06 22:50:01    403
2019-10-06 22:55:02    403
2019-10-06 23:00:01    403
2019-10-06 23:05:01    403
2019-10-06 23:10:02    403
2019-10-06 23:15:01    403
2019-10-06 23:20:01    403
2019-10-06 23:25:02    403
2019-10-06 23:30:01    403
2019-10-06 23:35:02    403
2019-10-06 23:40:01    403
2019-10-06 23:45:02    403
2019-10-06 23:50:02    403
2019-10-06 23:55:01    403
2019-10-07 00:00:02    403
2019-10-07 00:05:01    403
2019-10-07 00:10:02    403
2019-10-07 00:15:01    403
2019-10-07 00:20:02    403
2019-10-07 00:25:01    403
2019-10-07 00:30:02    403
2019-10-06 22:35:01    403
2019-10-07 00:35:01    403
2019-10-06 21:20:02    403
2019-10-06 21:25:02    403
2019-10-06 20:45:02    403
2019-10-06 20:50:01    403
2019-10-06 20:55:01    403
2019-10-06 21:00:01    403
2019-10-06 21:05:01    403
2019-10-06 21:10:01    403
2019-10-06 21:15:02    403
2019-10-06 22:20:02    403
2019-10-06 21:30:01    403
2019-10-06 21:35:01    403
2019-10-06 21:40:02    403
2019-10-06 21:45:02    403
2019-10-06 21:50:01    403
2019-10-06 21:55:01    403
2019-10-06 22:00:02    403
2019-10-06 22:05:02    403
2019-10-06 22:10:01    403
2019-10-06 22:15:01    403
2019-10-06 20:30:02    386
2019-10-06 20:20:02    386
2019-10-06 20:25:02    386
2019-10-06 20:15:02    386
2019-10-06 20:35:02    386
2019-10-06 20:40:02    386
2019-10-06 20:10:02    386
Name: count, dtype: int64
```

```
Name: count, dtype: int64
...
territoryName
Território Nacional    1134
Braga                  1134
Porto                  1134
Leiria                 1134
Madeira                1080
Coimbra                1080
Aveiro                 1080
Lisboa                 1080
Viseu                  1080
Viana do Castelo       1080
Setúbal                1080
Santarém               1026
Évora                  1026
Faro                   1026
Guarda                 1026
Bragança                972
Castelo Branco          972
Beja                    918
Portalegre              918
Vila Real               864
Açores                  799
Name: count, dtype: int64
...
totalMandates
0      5055
3      2949
2      1755
9      1726
6      1419
       ...
129      21
132      21
144      21
149      21
17       20
Name: count, Length: 62, dtype: int64
```

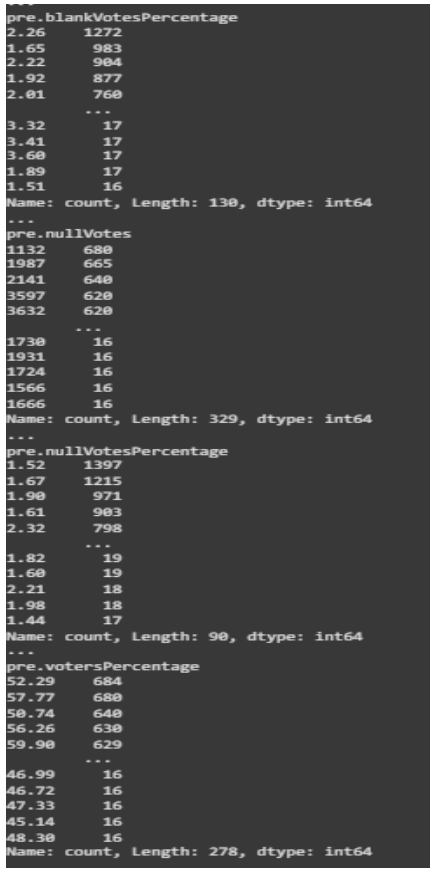
```
Name: count, Length: 62, dtype: int64
...
availableMandates
0      9686
3      1873
6      1201
9      1150
4      1089
       ...
144      21
28       21
30       21
42       21
44       20
Name: count, Length: 69, dtype: int64
...
numParishes
69      1944
3092    1134
243     1134
110     1134
347     1134
155     1080
147     1080
134     1080
54      1080
55      1080
208     1080
277     1080
67      1026
242     1026
141     1026
226      972
120      972
75       918
197      864
156      799
Name: count, dtype: int64
...
numParishesApproved
69     1220
242     728
75      680
54      679
208     640
       ...
188      16
170      16
180      16
171      16
83       16
Name: count, Length: 219, dtype: int64
```

```
...
blankVotes
1371    716
2026    665
3617    640
697     620
1680    608
       ...
1524     16
1722     16
1870     16
1606     16
2241     16
Name: count, Length: 329, dtype: int64
...
blankVotesPercentage
2.13    875
2.27    859
2.97    798
0.54    740
2.64    686
       ...
5.15     17
5.46     17
5.01     17
4.87     17
1.98     16
Name: count, Length: 146, dtype: int64
...
nullVotes
1094    680
2057    665
2292    620
1047    608
988     595
       ...
1887     16
1985     16
2361     16
618      16
1719     16
Name: count, Length: 331, dtype: int64
...
nullVotesPercentage
1.77    1117
1.63     824
1.93     791
1.78     728
2.30     701
        ...
2.29      18
2.67      18
1.35      17
2.46      16
2.70      16
Name: count, Length: 107, dtype: int64
...
```

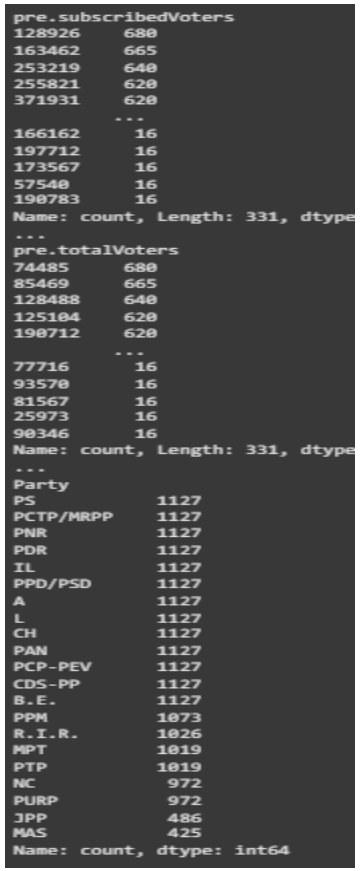
Broj jedinstvenih vrijednosti po stupcima #2



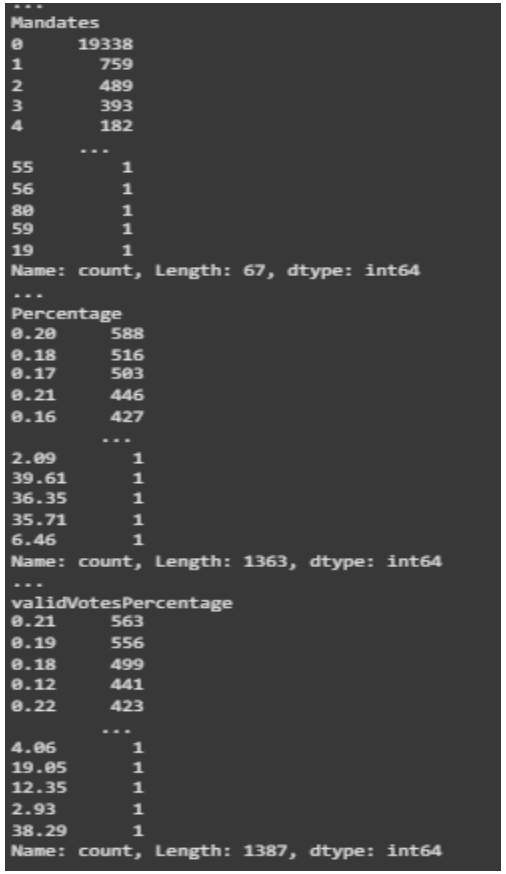
```
votersPercentage
52.26    714
50.58    665
50.60    640
50.29    620
55.09    612
        ...
45.74     16
44.20     16
44.82     16
44.21     16
44.28     16
Name: count, Length: 282, dtype: int64
...
subscribedVoters
122987    680
151535    665
240917    640
258144    620
136696    608
         ...
161553     16
216808     16
186681     16
53405      16
181546     16
Name: count, Length: 335, dtype: int64
...
totalVoters
64269     680
76649     665
129821    620
74026     608
51212     595
         ...
73425      16
77266      16
81364      16
22931      16
84261      16
Name: count, Length: 336, dtype: int64
...
pre.blankVotes
1230    680
1689    665
3228    640
4235    620
1018    620
       ...
1587     16
1377     16
1534     16
393      16
1451     16
Name: count, Length: 323, dtype: int64
...
```



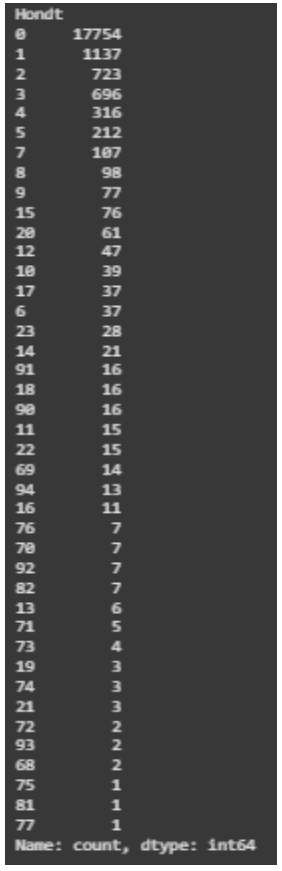
```
...
pre.blankVotesPercentage
2.26    1272
1.65     983
2.22     904
1.92     877
2.01     760
        ...
3.32      17
3.41      17
3.60      17
1.89      17
1.51      16
Name: count, Length: 130, dtype: int64
...
pre.nullVotes
1132    680
1987    665
2141    640
3597    620
3632    620
       ...
1730     16
1931     16
1724     16
1566     16
1666     16
Name: count, Length: 329, dtype: int64
...
pre.nullVotesPercentage
1.52    1397
1.67    1215
1.90     971
1.61     903
2.32     798
        ...
1.82      19
1.60      19
2.21      18
1.98      18
1.44      17
Name: count, Length: 90, dtype: int64
...
pre.votersPercentage
52.29    684
57.77    680
50.74    640
56.26    630
59.90    629
        ...
46.99     16
46.72     16
47.33     16
45.14     16
48.30     16
Name: count, Length: 278, dtype: int64
```



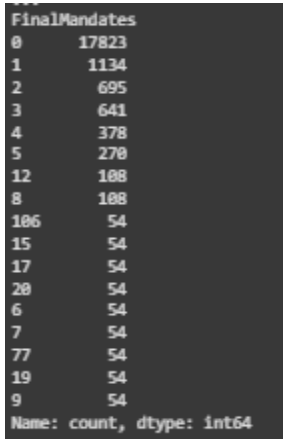
```
pre.subscribedVoters
128926    680
163462    665
253219    640
255821    620
371931    620
         ...
166162     16
197712     16
173567     16
57540      16
190783     16
Name: count, Length: 331, dtype
...
pre.totalVoters
74485     680
85469     665
128488    640
125104    620
190712    620
         ...
77716      16
93570      16
81567      16
25973      16
90346      16
Name: count, Length: 331, dtype
...
Party
PS           1127
PCTP/MRPP    1127
PNR          1127
PDR          1127
IL           1127
PPD/PSD      1127
A            1127
L            1127
CH           1127
PAN          1127
PCP-PEV      1127
CDS-PP       1127
B.E.         1127
PPM          1073
R.I.R.       1026
MPT          1019
PTP          1019
NC            972
PURP          972
JPP           486
MAS           425
Name: count, dtype: int64
```



```
...
Mandates
0     19338
1       759
2       489
3       393
4       182
       ...
55        1
56        1
80        1
59        1
19        1
Name: count, Length: 67, dtype: int64
...
Percentage
0.20     588
0.18     516
0.17     503
0.21     446
0.16     427
        ...
2.09       1
39.61      1
36.35      1
35.71      1
6.46       1
Name: count, Length: 1363, dtype: int64
...
validVotesPercentage
0.21     563
0.19     556
0.18     499
0.12     441
0.22     423
        ...
4.06       1
19.05      1
12.35      1
2.93       1
38.29      1
Name: count, Length: 1387, dtype: int64
```



```
Hondt
0     17754
1      1137
2       723
3       696
4       316
5       212
7       107
8        98
9        77
15       76
20       61
12       47
10       39
17       37
6        37
23       28
14       21
91       16
18       16
90       16
11       15
22       15
69       14
94       13
16       11
76        7
70        7
92        7
82        7
13        6
71        5
73        4
19        3
74        3
21        3
72        2
93        2
68        2
75        1
81        1
77        1
Name: count, dtype: int64
```



```
...
FinalMandates
0     17823
1      1134
2       695
3       641
4       378
5       270
12      108
8       108
106      54
15       54
17       54
20       54
6        54
7        54
77       54
19       54
9        54
Name: count, dtype: int64
```

Dimenzije dataseta

- Neke od dimenzija po kojima bi se ovaj dataset mogao promatrati su:
- Vrijeme
- Teritorj
- Stranke
- Mandati
- Glasovi
- ...