



ISTRAŽIVANJE SAOBRAĆAJNIH NESREĆA U FRANCUSKOJ

Istraživanje Podataka 1
Matematički fakultet
Aleksa Knežević i Darko Nešković
mi16009@alas.matf.bg.ac.rs
mi16208@alas.matf.bg.ac.rs

Sadržaj

1. Uvod

1.1 Motivacija

1.2 Skup podataka o saobraćajnim nesrećama

2. Priprema podataka

2.1 Python pretprocesiranje

2.2 SPSS modeler pretprocesiranje

3. Primene algoritama klasterovanja

4. Diskusija I zanimljivosti

Uvod

Motivacija nam je bila da pronađemo veze praznika i broja i osobina saobraćajnih nesreća koje su se dešavale u periodu između 2005. i 2016.

Takođe smo želeli da analiziramo neke zanimljive statistike na datom skupu podataka.

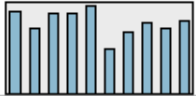
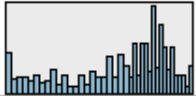
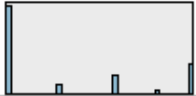
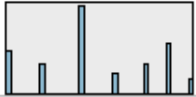
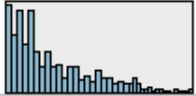


Link ka podacima:

<https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016>

Uvod

Broj slogova: 80.000

Broj atributa: 20

Field	Sample Graph	Measurement	Min	Max	Sum	Range	Mean	Mean Std. Err.	Std. Dev	Variance	Skewness
praznik		Categorical	--	--	--	--	--	--	--	--	--
sat_minut		Continuous	1	2359	108558236	2358	1371.031	2.186	615.154	378413.984	-0.638
osvetljenje		Continuous	1	5	166787	4	2.106	0.006	1.595	2.543	1.005
tip_sudara		Continuous	1	7	287054	6	3.625	0.006	1.802	3.249	0.240
postanski_broj_opstine		Continuous	1	922	15107670	921	190.802	0.615	173.114	29968.361	1.158
uslovi_na_putu		Continuous	0	9	101055	9	1.276	0.003	0.961	0.924	5.576
stepen_povrede		Continuous	1	4	195719	3	2.472	0.005	1.324	1.752	-0.046

Uvid u podatke i značajne slogove

Priprema Podataka

PYTHON

- **Regex za delimiter “,” u Characteristics.csv**
- **Izbacivanje svih slogova koji poseduju NULL vrednosti**
- **Binarizacija kategoričkih atributa**

Priprema podataka

SPSS modeler

- Spajanje svih tabela, osim tabele holidays po atributu id_nesreće
- Spajanje atributa dan/mesec/godina u atribut datum
- Povezivanje preko tog atributa sa tabelom praznik
- Binarizacija kategoričkih podataka

Primene algoritama klasterovanja

K-sredina

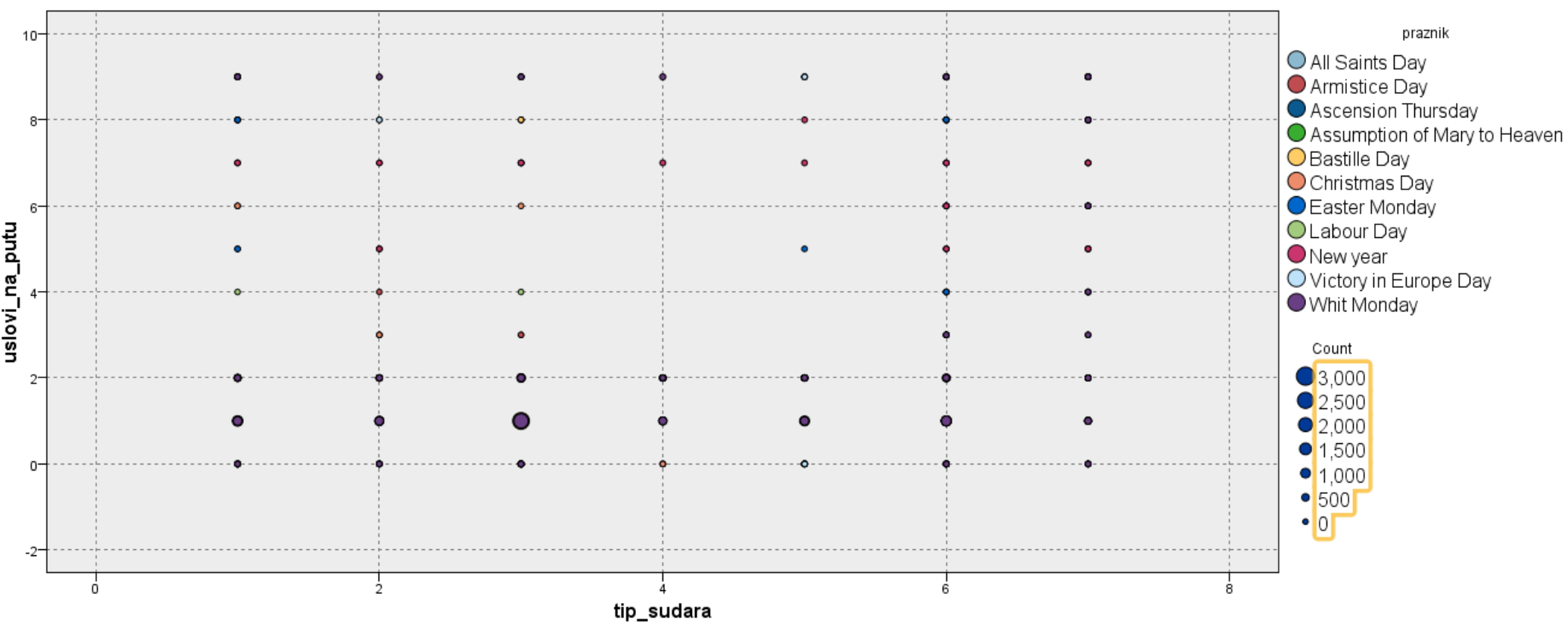
Model Summary

Algorithm	K-Means
Inputs	22
Clusters	60

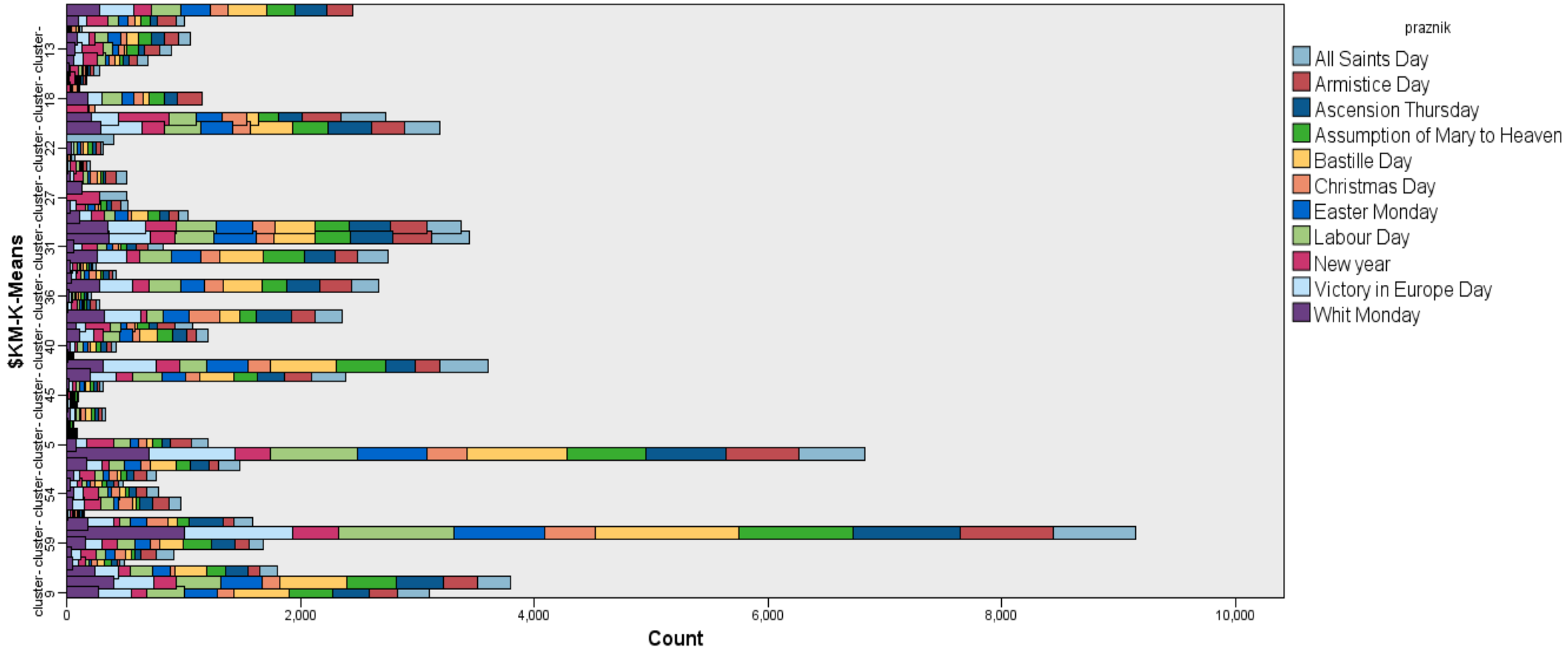
Average Silhouette= 0.7



Primene algoritama klasterovanja



Primene algoritama klasterovanja

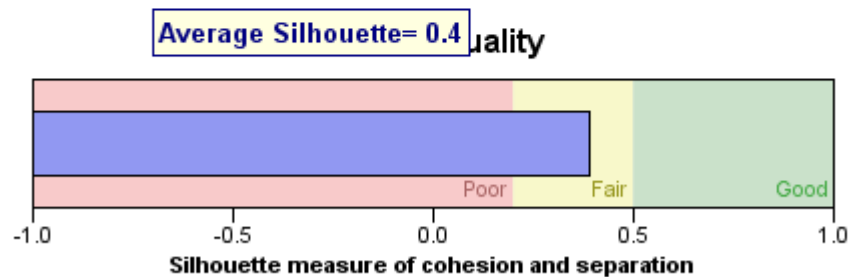


Primene algoritama klasterovanja

Kohonen

Model Summary

Algorithm	Kohonen
Inputs	22
Clusters	20



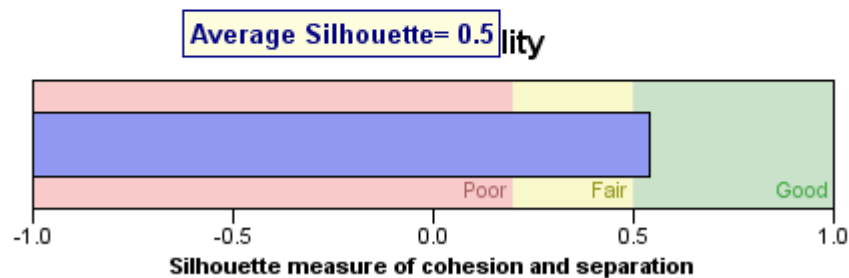
Pre PCA (principal component analysis)

Primene algoritama klasterovanja

Kohonen

Model Summary

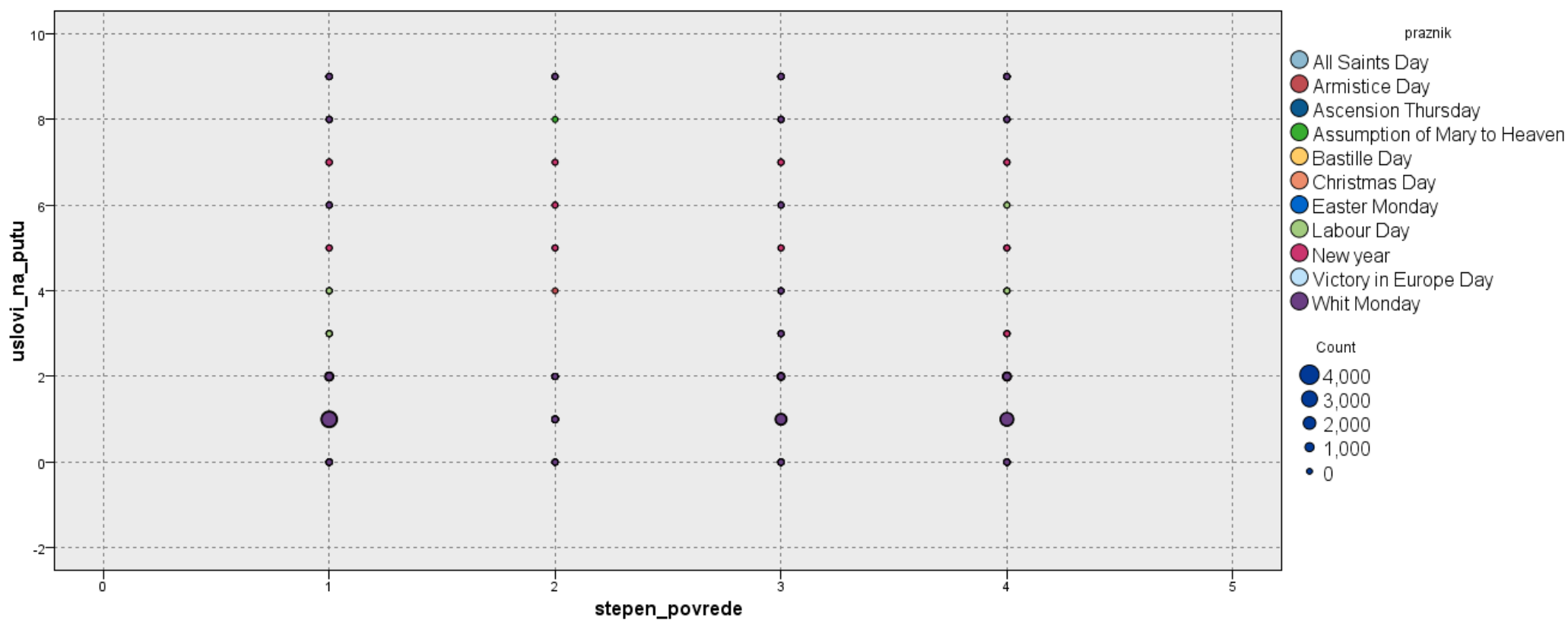
Algorithm	Kohonen
Inputs	25
Clusters	30



Posle PCA

Primene algoritama klasterovanja

Kohonen

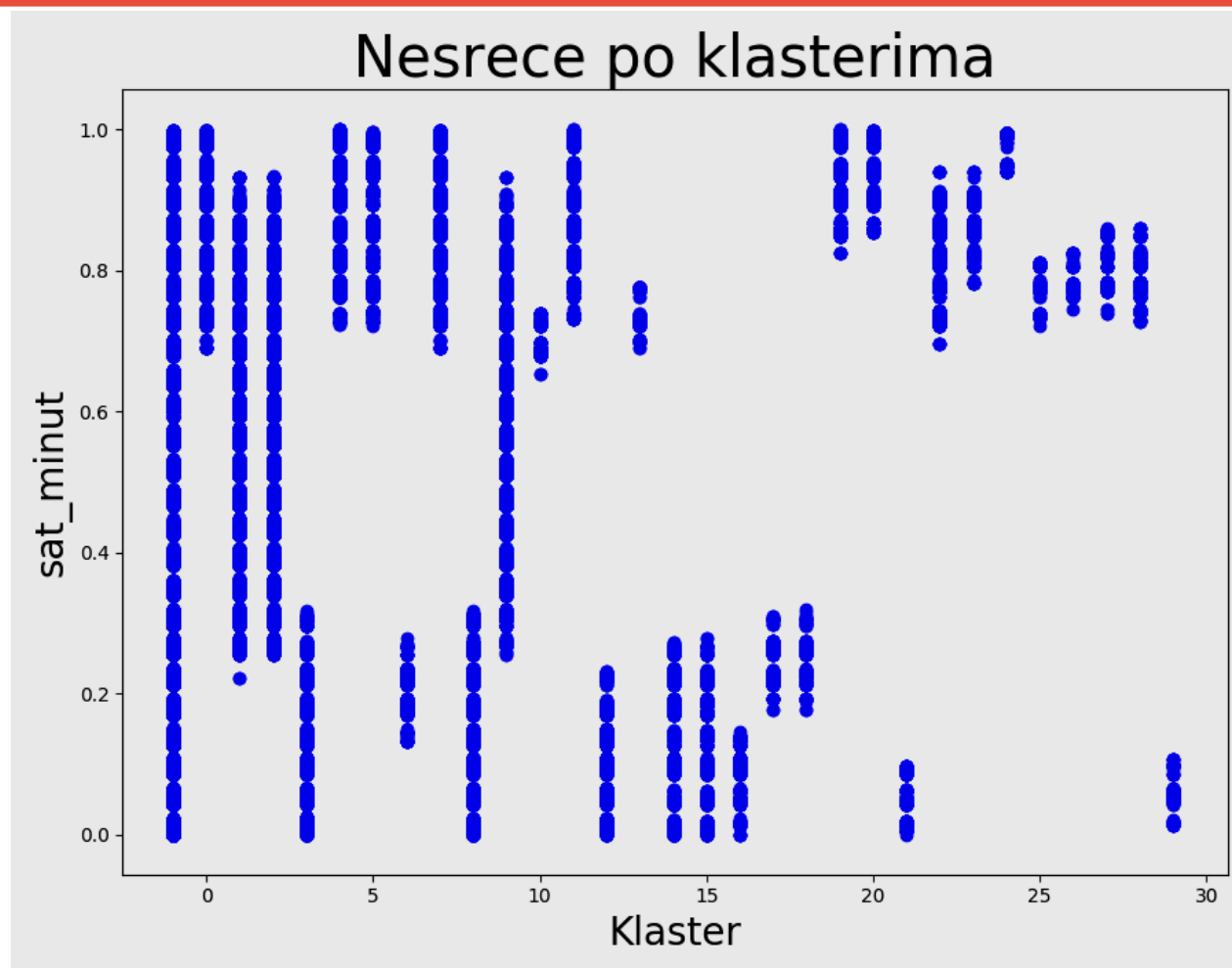


Primene algoritama klasterovanja

DBSCAN

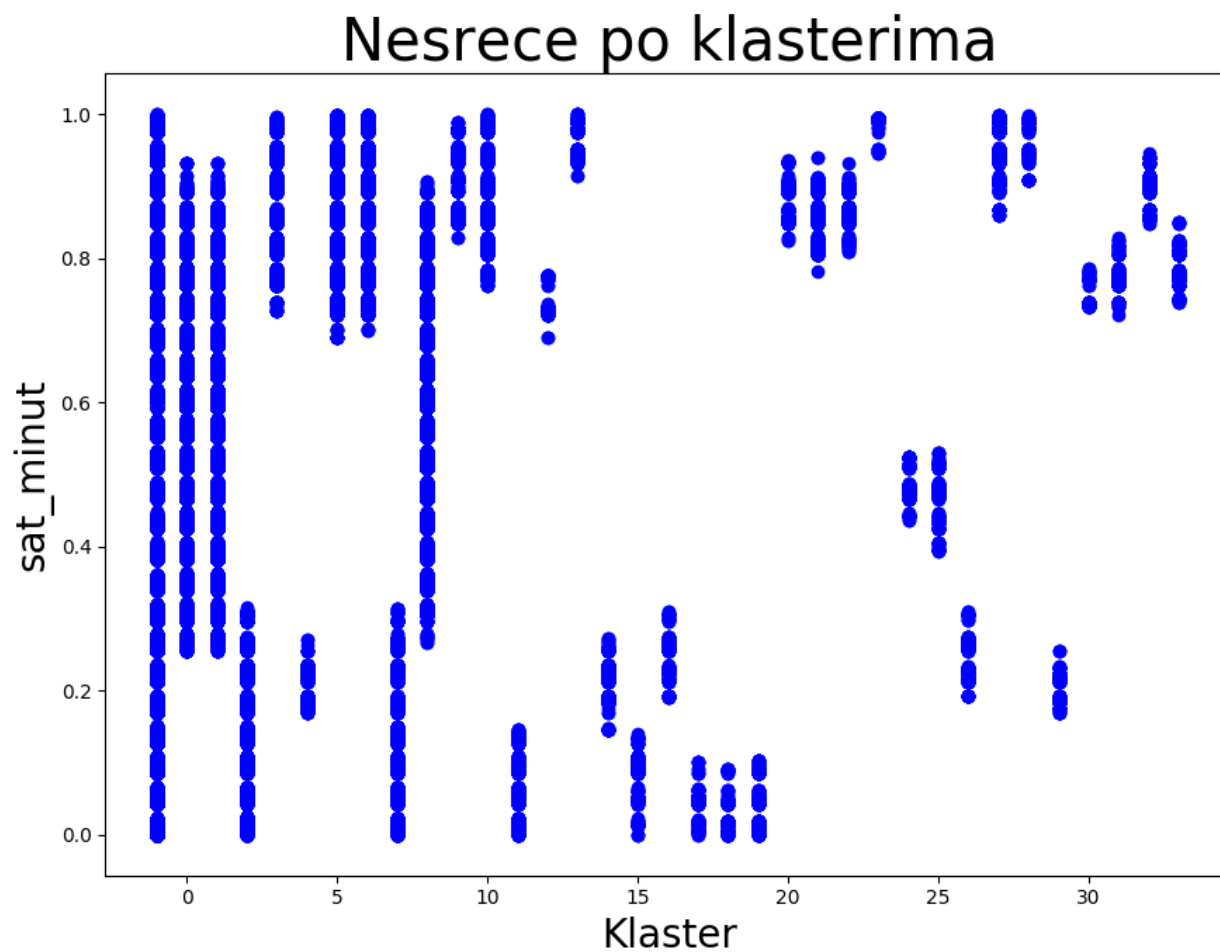
Broj klastera	Senka Koeficijent	Epsilon
31	0.75312	0.048
35	0.697891	0.046
39	0.701062	0.044

Primene algoritama klasterovanja



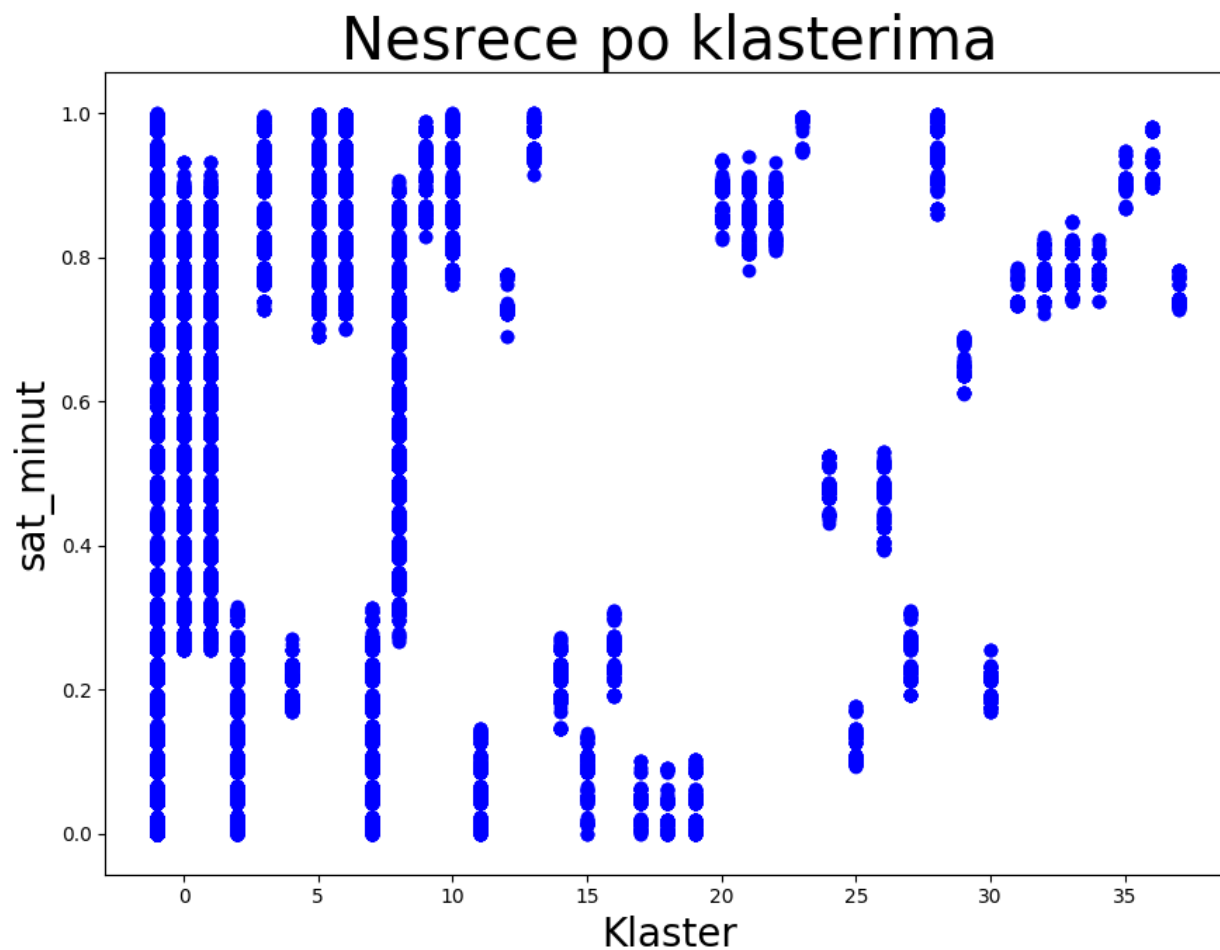
Epsilon: 0.048

Primene algoritama klasterovanja



Epsilon: 0.046

Primene algoritama klasterovanja



Epsilon: 0.044

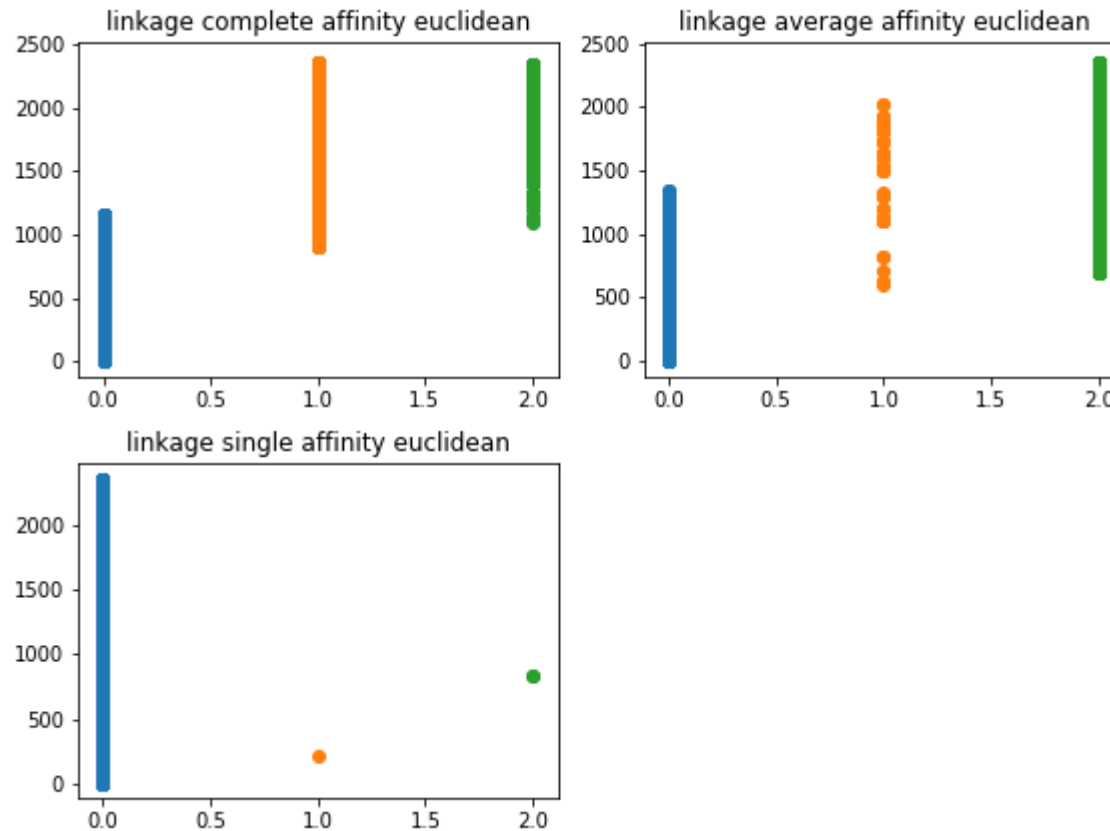
Primene algoritama klasterovanja

Hijerarhijsko Klasterovanje

Broj klastera	Senka koeficijent	rastojanje	sklonost
3	0.4741	euklidsko	pojedinačna
3	0.4303	euklidsko	prosečna
3	0.4090	euklidsko	kompletna
3	0.4665	menhetn	pojedinačna
3	0.3967	menhetn	prosečna
3	0.4298	menhetn	kompletna

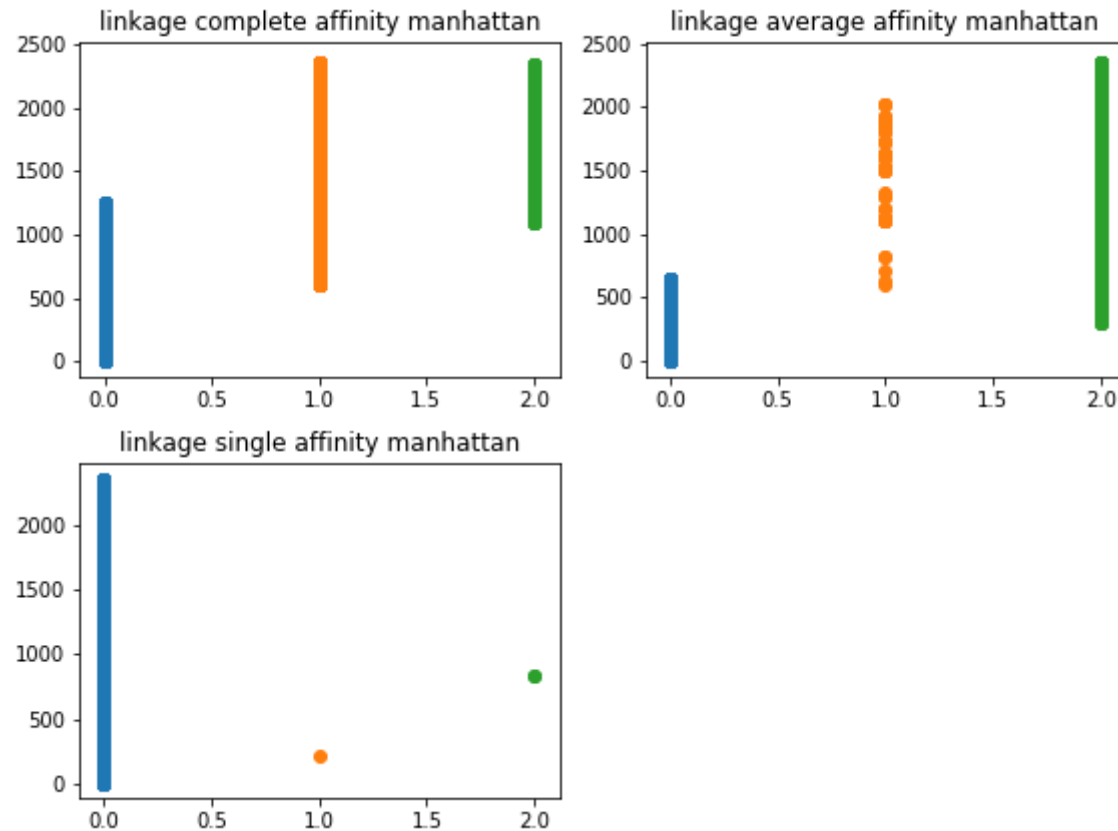
Primene algoritama klasterovanja

Euklidsko rastojanje

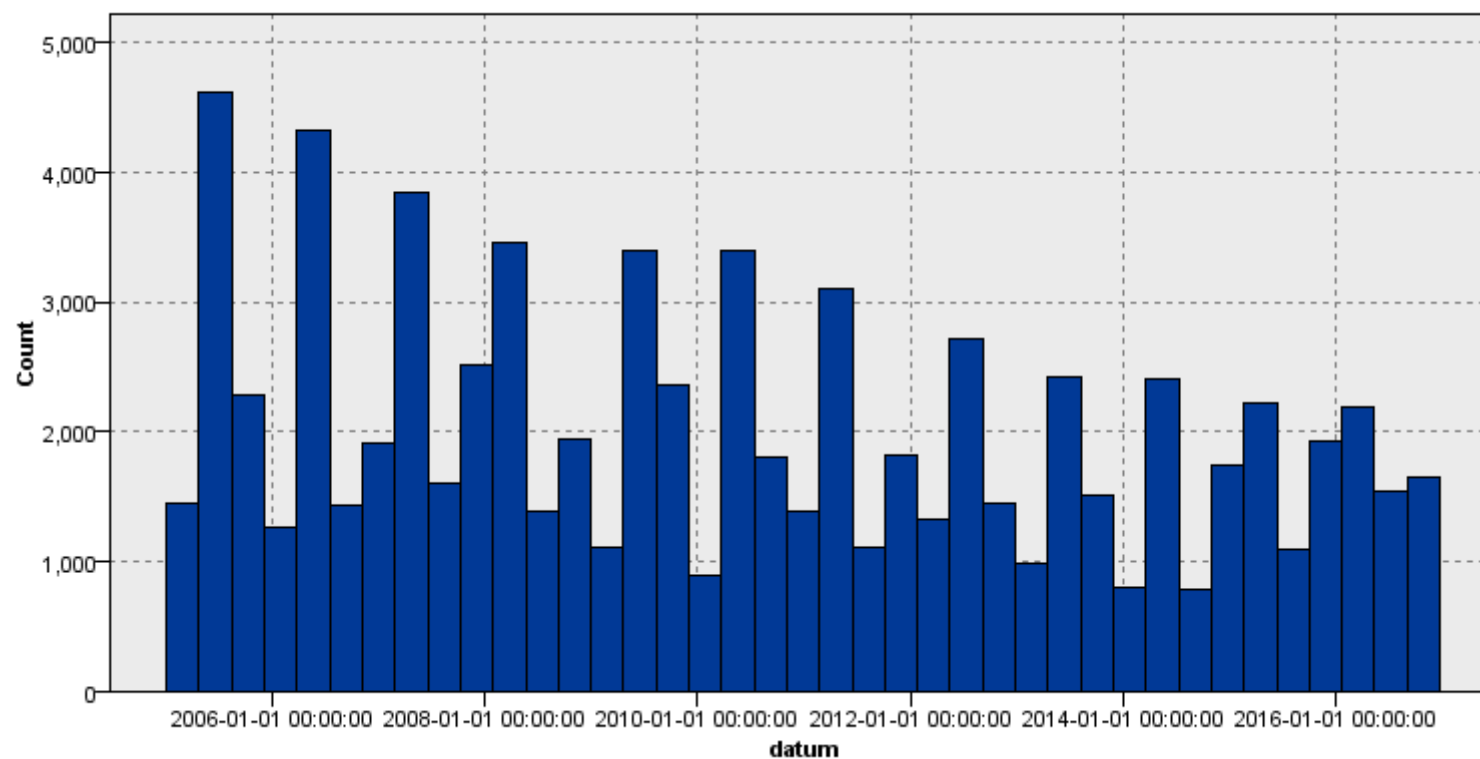


Primene algoritama klasterovanja












Menhetn rastojanje



Zanimljivosti



Zanimljivosti

Value	Proportion π	%	Count
Bastille Day		10.7	8476
All Saints Day		9.94	7869
Victory in Europe Day		9.85	7802
Ascension Thursday		9.71	7686
Labour Day		9.55	7562
Whit Monday		9.54	7555
Armistice Day		9.24	7319
Assumption of Mary to Heaven		9.16	7249
Easter Monday		8.33	6596
New year		7.91	6261
Christmas Day		6.07	4805