Predefinisani projekat 2

1. Prvi projekat

Iz zadatog skupa podataka (*credit_card_data.csv*) potrebno je grupisati korisnike koji imaju slično ponašanje. Odrediti optimalan broj grupa i opisati ih.

U *preprocesing_data* funkciji odradili smo preprocesiranje podataka. Prvo smo pokušali da odbacimo redove u kojima nedostaje vrednost, pa smo odustali od toga jer smo izgubili dosta podataka. Odlučili smo se da prazna polja popunimo prosečnim vrednostima. Pomoću *describe* možemo da vidimo rasprostranjenost podataka u svakoj koloni našeg csv fajla (slike 1 i 2).

	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	\
count	8950.000000	8950.000000	8950.000000	8950.000000	
mean	1564.474828	0.877271	1003.204834	592.437371	
std	2081.531879	0.236904	2136.634782	1659.887917	
min	0.000000	0.000000	0.000000	0.000000	
25%	128.281915	0.888889	39.635000	0.000000	
50%	873.385231	1.000000	361.280000	38.000000	
75%	2054.140036	1.000000	1110.130000	577.405000	
max	19043.138560	1.000000	49039.570000	40761.250000	

	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	\
count	8950.000000	8950.000000	8950.000000	
mean	411.067645	978.871112	0.490351	
std	904.338115	2097.163877	0.401371	
min	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.083333	
50%	89.000000	0.000000	0.500000	
75%	468.637500	1113.821139	0.916667	
max	22500.000000	47137.211760	1.000000	

	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY	\
count	8950.000000	8950.000000	
mean	0.202458	0.364437	
std	0.298336	0.397448	
min	0.000000	0.000000	
25%	0.000000	0.000000	
50%	0.083333	0.166667	
75%	0.300000	0.750000	
max	1.000000	1.000000	

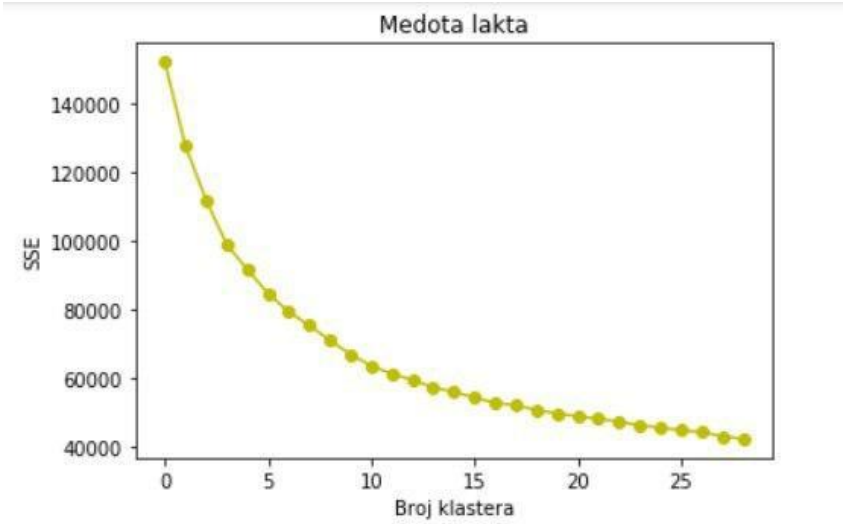
Slika 1.

	CASH_ADVANCE_FREQUENCY	CASH_ADVANCE_TRX	PURCHASES_TRX	CREDIT_LIMIT	\
count	8950.000000	8950.000000	8950.000000	8950.000000	
mean	0.135144	3.248827	14.709832	4494.282473	
std	0.200121	6.824647	24.857649	3638.646702	
min	0.000000	0.000000	0.000000	50.000000	
25%	0.000000	0.000000	1.000000	1600.000000	
50%	0.000000	0.000000	7.000000	3000.000000	
75%	0.222222	4.000000	17.000000	6500.000000	
max	1.500000	123.000000	358.000000	30000.000000	

	PAYMENTS	MINIMUM_PAYMENTS	PRC_FULL_PAYMENT	TENURE
count	8950.000000	8950.000000	8950.000000	8950.000000
mean	1733.143852	844.906767	0.153715	11.517318
std	2895.063757	2332.792322	0.292499	1.338331
min	0.000000	0.019163	0.000000	6.000000
25%	383.276166	170.857654	0.000000	12.000000
50%	856.901546	312.343947	0.000000	12.000000
75%	1901.134317	788.713501	0.142857	12.000000
max	50721.483360	76406.207520	1.000000	12.000000

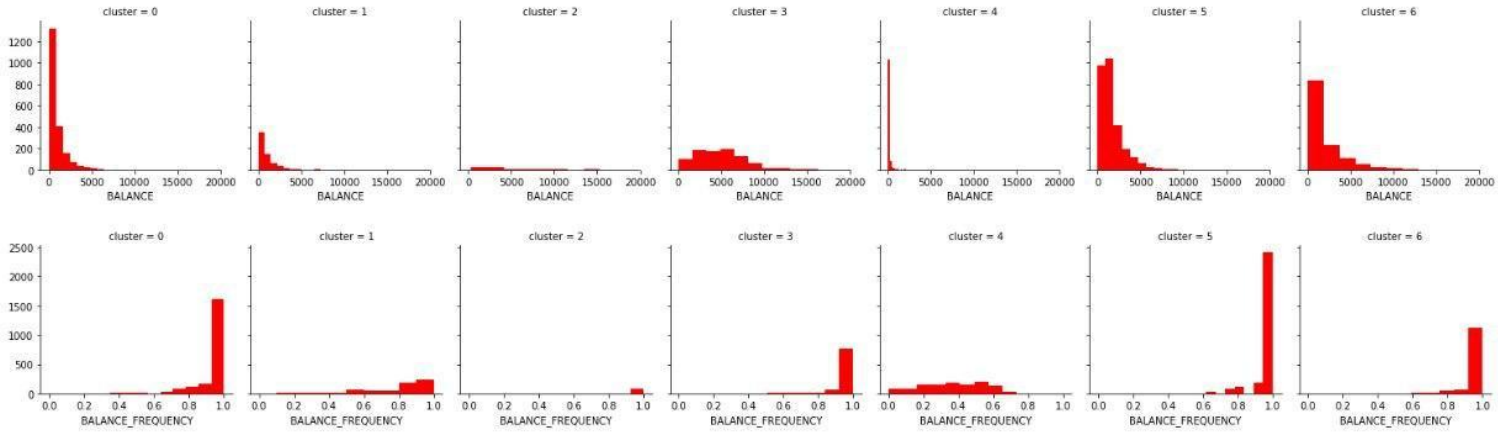
Slika 2.

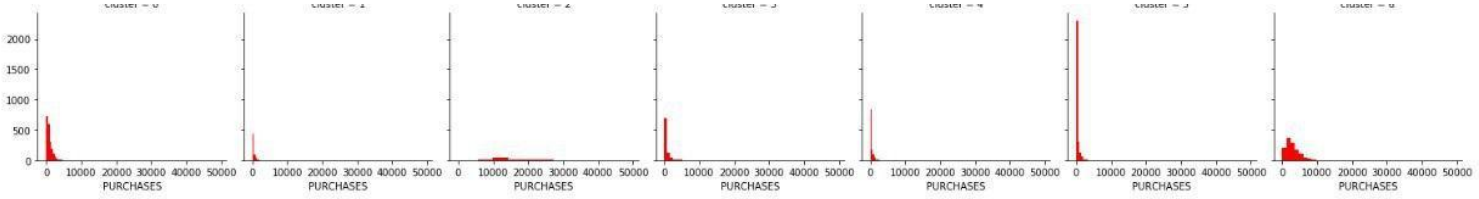
U *normalize_data* funkciji smo normalizovali podatke pomoću StandardScalera. Nakon toga smo u *vizualization* funkciji primenili metodu lakta (Slika 3) kako bismo odredili broj klastera.

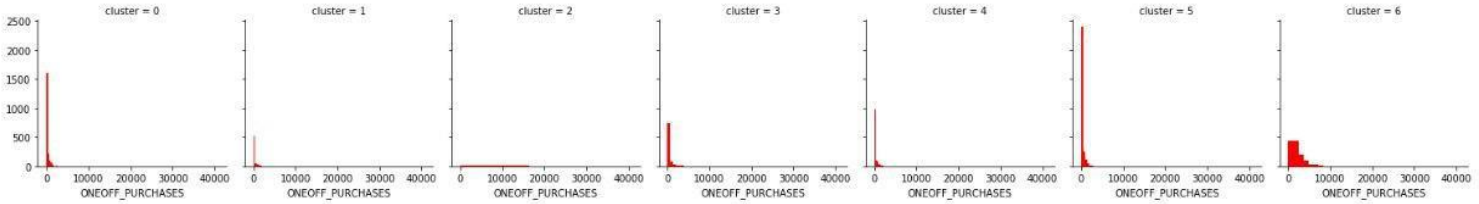


Slika 3.

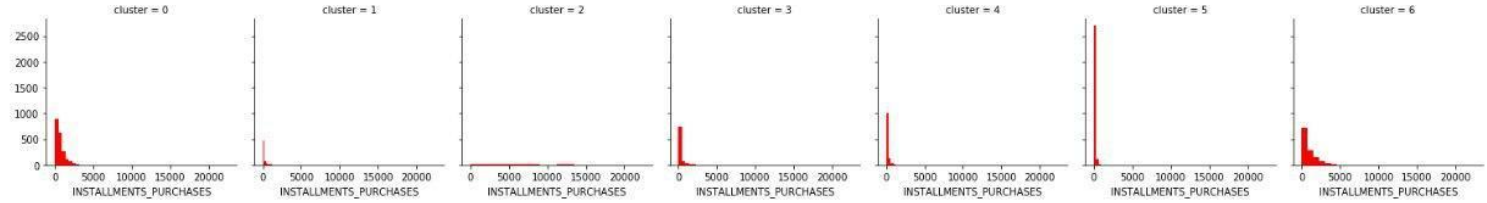
Kako možemo da vidimo pad krive između 5 i 10, uzeli smo broj 7 za broj klastera. Nakon obrađenih sedam klastera iscrtali smo ih u histogramima kako bismo lakše opisali svaki od njih ponaosob. Na sledećim slikama prikazane su vrednosti svih kolona za svaki klaster.

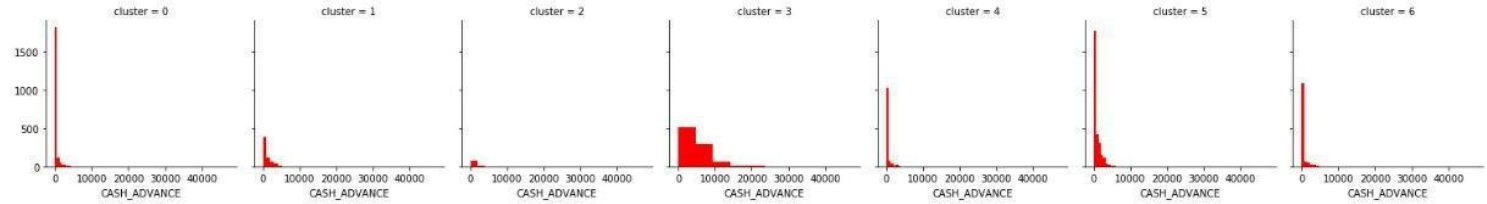


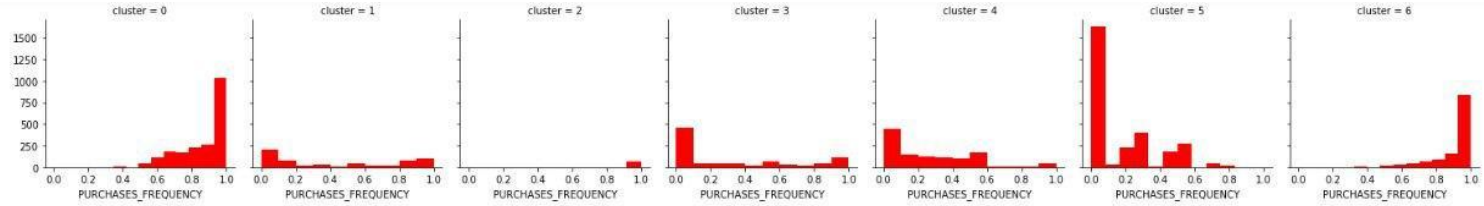

PURCHASES

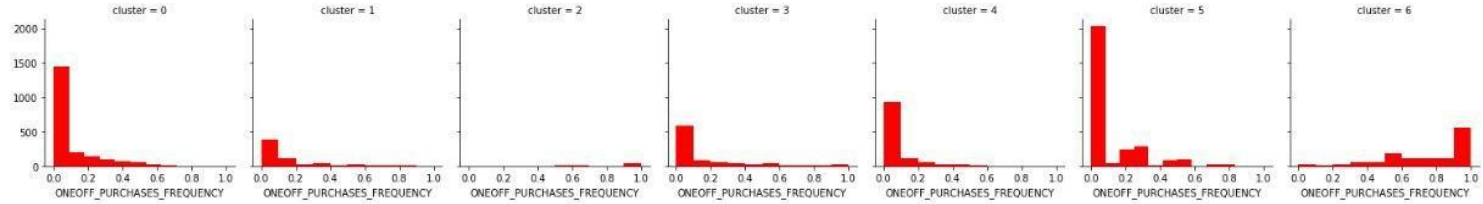

cluster = 0
cluster = 1
cluster = 2
cluster = 3
cluster = 4
cluster = 5
cluster = 6
ONEOFF_PURCHASES

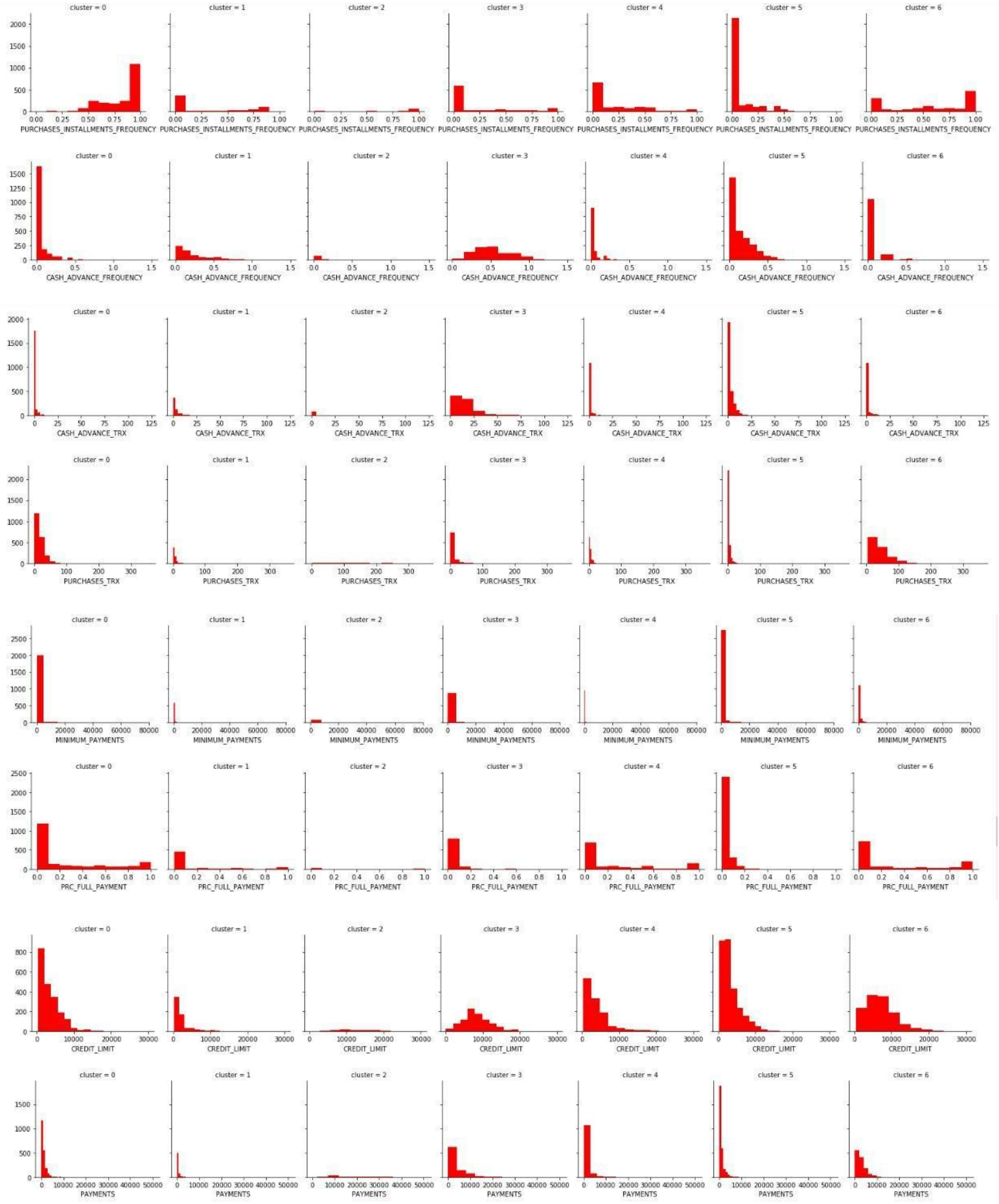

cluster = 0
cluster = 1
cluster = 2
cluster = 3
cluster = 4
cluster = 5
cluster = 6

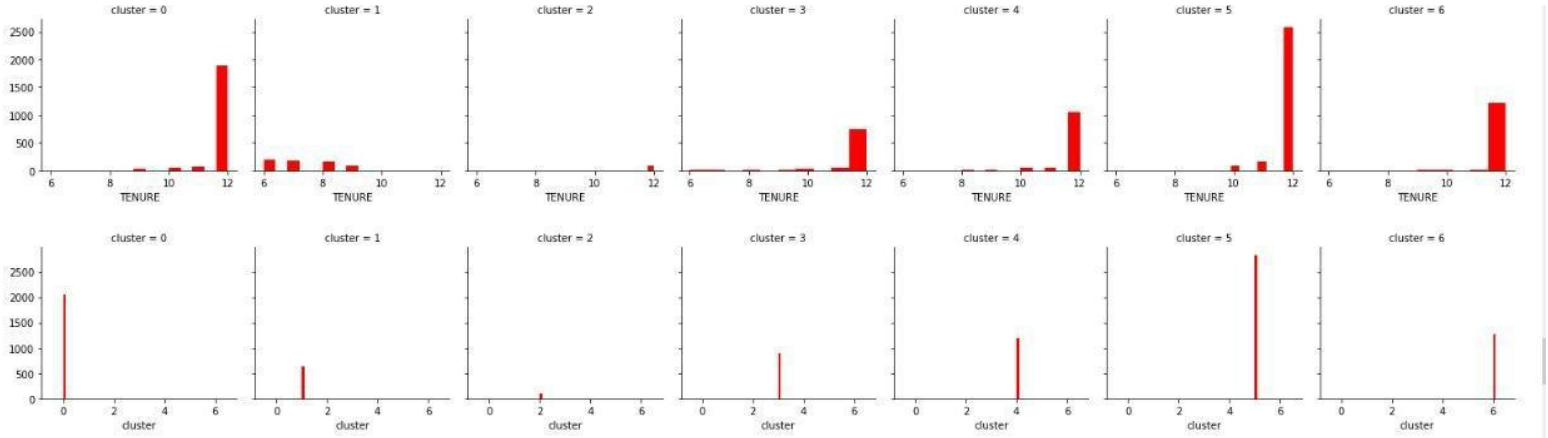

cluster = 0
cluster = 1
cluster = 2
cluster = 3
cluster = 4
cluster = 5
cluster = 6
INSTALLMENTS_PURCHASES


cluster = 0
cluster = 1
cluster = 2
cluster = 3
cluster = 4
cluster = 5
cluster = 6
CASH_ADVANCE


cluster = 0
cluster = 1
cluster = 2
cluster = 3
cluster = 4
cluster = 5
cluster = 6
PURCHASES_FREQUENCY


cluster = 0
cluster = 1
cluster = 2
cluster = 3
cluster = 4
cluster = 5
cluster = 6
ONEOFF_PURCHASES_FREQUENCY

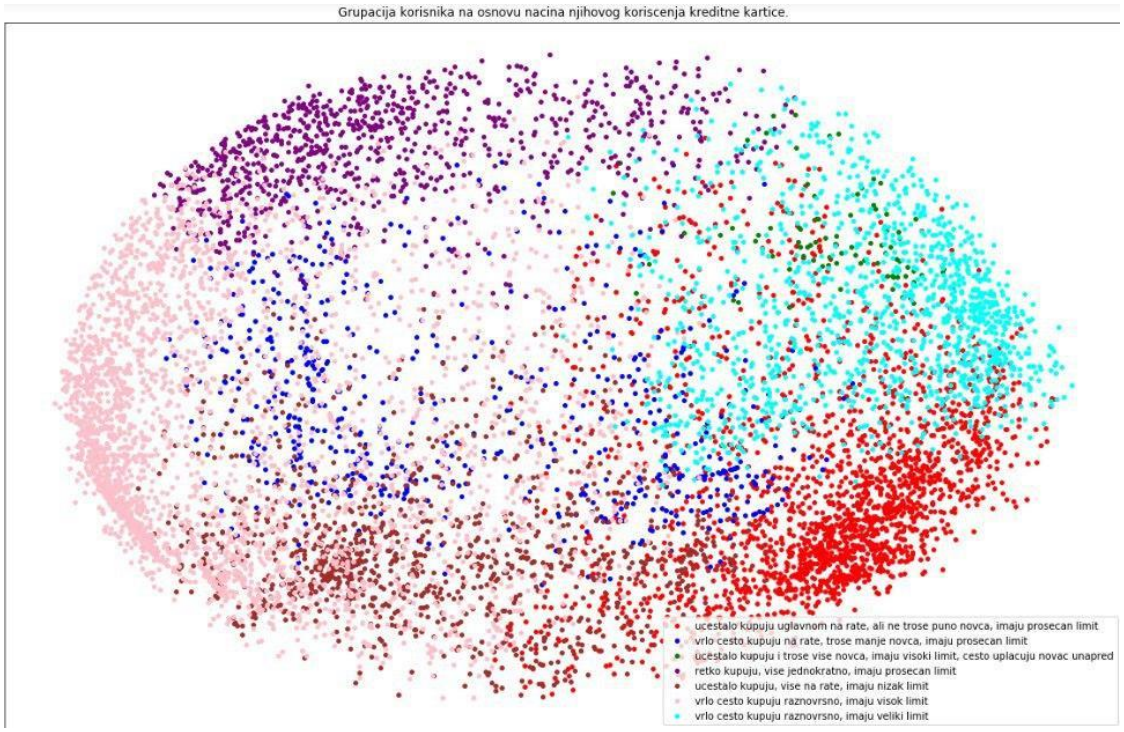

cluster = 0
cluster = 1
cluster = 2
cluster = 3
cluster = 4
cluster = 5
cluster = 6
PURCHASES_INSTALLMENTS_FREQUENCY
CASH_ADVANCE_FREQUENCY
CASH_ADVANCE_TRX
PURCHASES_TRX
MINIMUM_PAYMENTS
PRC_FULL_PAYMENT
CREDIT_LIMIT
PAYMENTS



Možemo da vidimo sledeću raspodelu među klasterima:

1) Klaster 0: korisnici koji kupuju uglavnom na rate ali ne troše puno novca i imaju prosečan limit na kreditnoj kartici
2) Klaster 1: čine korisnici koji vrlo često kupuju na rate, troše manje novca i imaju prosečan limit na kreditnoj kartici
3) Klaster 2: korisnici koji učestalo kupuju i troše puno novca, imaju visok limit na kreditnoj kartici i često uplaćuju novac unapred
4) Klaster 3: čine korisnici koji retko kupuju, više jednokratno i imaju prosečan limit na kreditnoj kartici
5) Klaster 4: korinsici koji učestalo kupuju više na rate, imaju nizak limit na kreditnoj kartici
6) Klaster 5: čine korisnici koji često kupuju raznovrsno i imaju visok limit na kreditnoj kartici
7) Klaster 6: korisnici koji vrlo često kupuju raznovrsno i imaju visok limit na kreditnoj kartici.

Za vizuelizaciju podataka koristili smo PCA algoritam. On se koristi za redukciju kod podataka koji imaju dosta dimenzija, kao što je slučaj kod nas. Zato smo pomoću PCA redukovali podatke u dve dimenzije, na slici 4 može se videti grupacija korisnika na osnovu korišćenja njihove kretine kartice.



Slika 4.