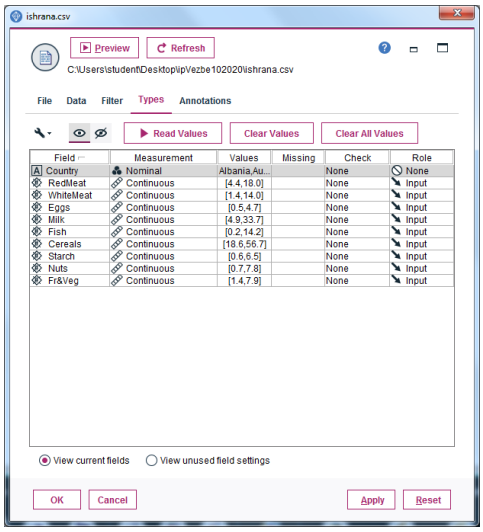
# Istraživanje podataka 1 - vežbe 10, 2020.

Primer 2: Izvršiti klasterovanje država na osnovu podataka o ishrani stanovništva primenom algoritma K-sredina u alatu IBM SPSS Modeler. Skup *ishrana.csv* sadrži podatak koliko su određene namirnice zastupljene u ishrani stanovništva nekih država. Vrednosti numeričkih atributa predstavljaju procente. Atributi skupa su:

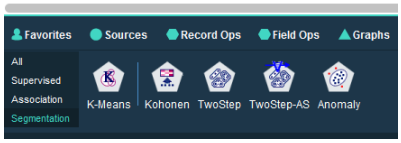
- *Country* - država
- *RedMeat* - crveno meso
- *Eggs* - jaja
- *Milk* - mleko
- *Fish* - riba
- *Cereals* - žitarice
- *Starch* - skrob
- *Nuts* - koštunjavo voće
- *Fr&Veg* - voće i povrće

U radnom toku **drzave_ishrana.str** se prvo učitava skup pomoću čvora *Var. File*. U odeljku *Types* klikom na dugme *Read Values* učitavaju se informacije o vrednostima koje se javljaju u atributima skupa. Atributima koji učestvuju u klasterovanju uloga (*Role*) se postavlja na *Input*. U ovom primeru za klasterovanje se koriste svi numerički atributi. Pošto svaka država ima jedinstveno ime, atribut *Country* nema značaj u klasterovanju, te se njegova uloga postavlja na *None*, tj. atribut neće biti korišćen pri klasterovanju (Slika 1).



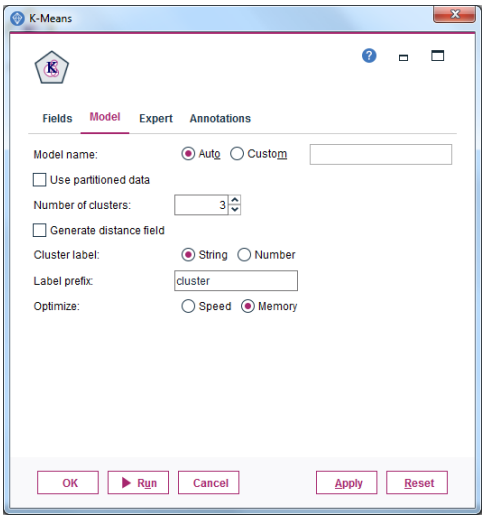

**Slika 1:** Učitavanju vrednosti u atributima skupa i dodela uloga atributima

Da bi se primenio algoritm K-sredina na skup, čvor sa skupom podataka povezuje se sa čvorom *K-means* (podsećanje kako: klik na čvor sa skupom, taster F2, klik na čvor *K-means*) (Slika 2).

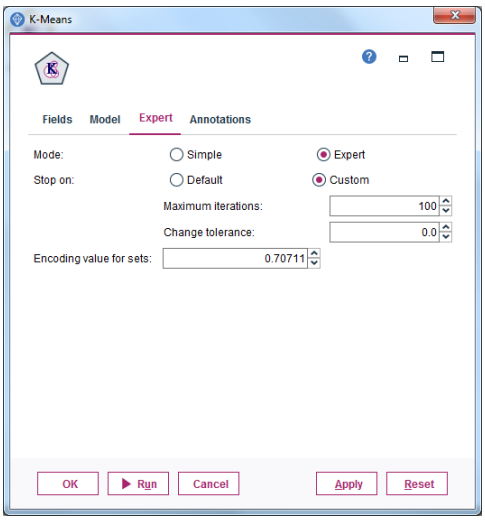

**Slika 2:** Izbor čvora *K-means*

Preko opcija dostupnih u čvoru *K-means*, u odeljku *Model* postavlja se, za početak, broj željenih klastera na 3. (Slika 3)



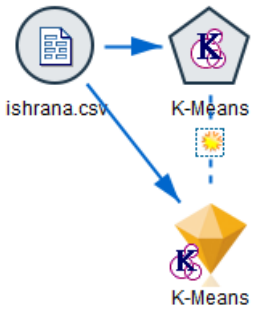

**Slika 3:** Postavljanje broja klastera u čvoru *K-means*

U odeljku *Expert*, maksimalan broj iteracija koje mogu biti izvršene u algoritmu K-sredina se povećava na 100 (Slika 4).




**Slika 4:** Postavljanje vrednosti za beoj iteracija u čvoru *K-means*

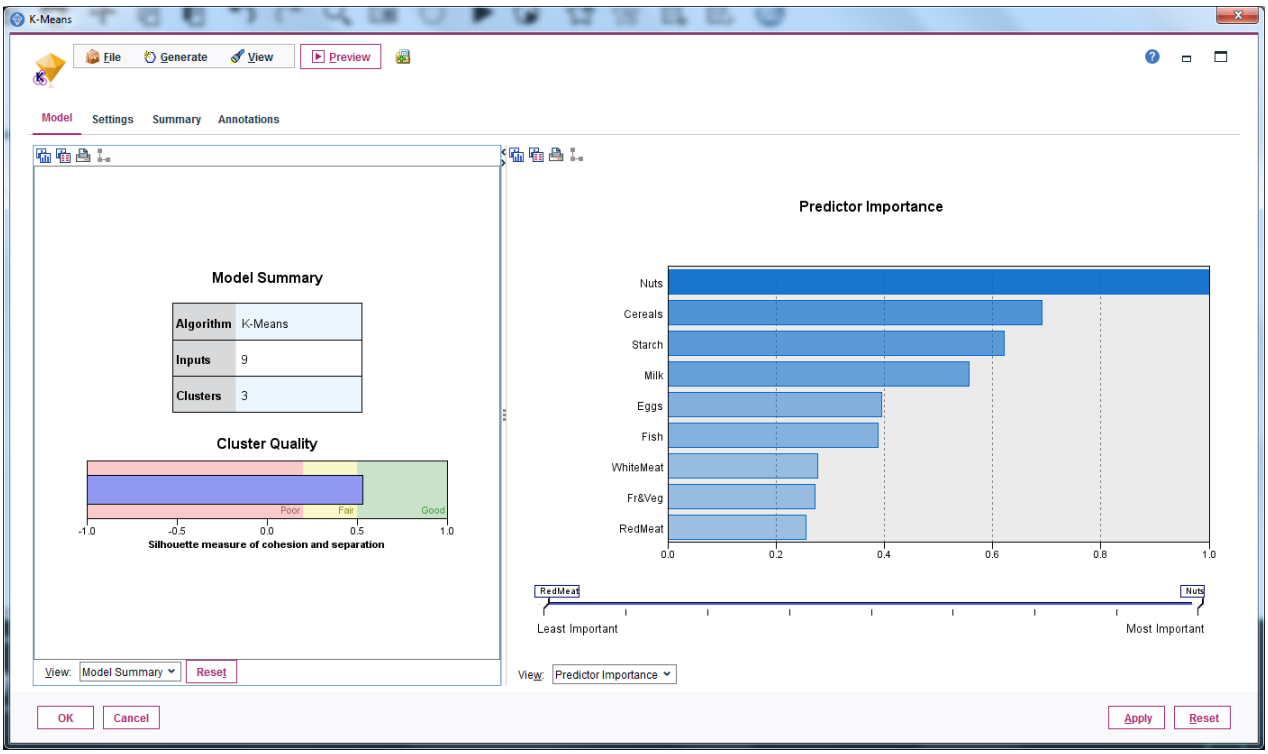
Izborom opcije *Run* pravi se model klasterovanja koji je u radnom toku prikazan čvorom u obliku dijamanta (Slika 5).




**Slika 5:** Radni tok sa napravljenim modelom klasterovanja

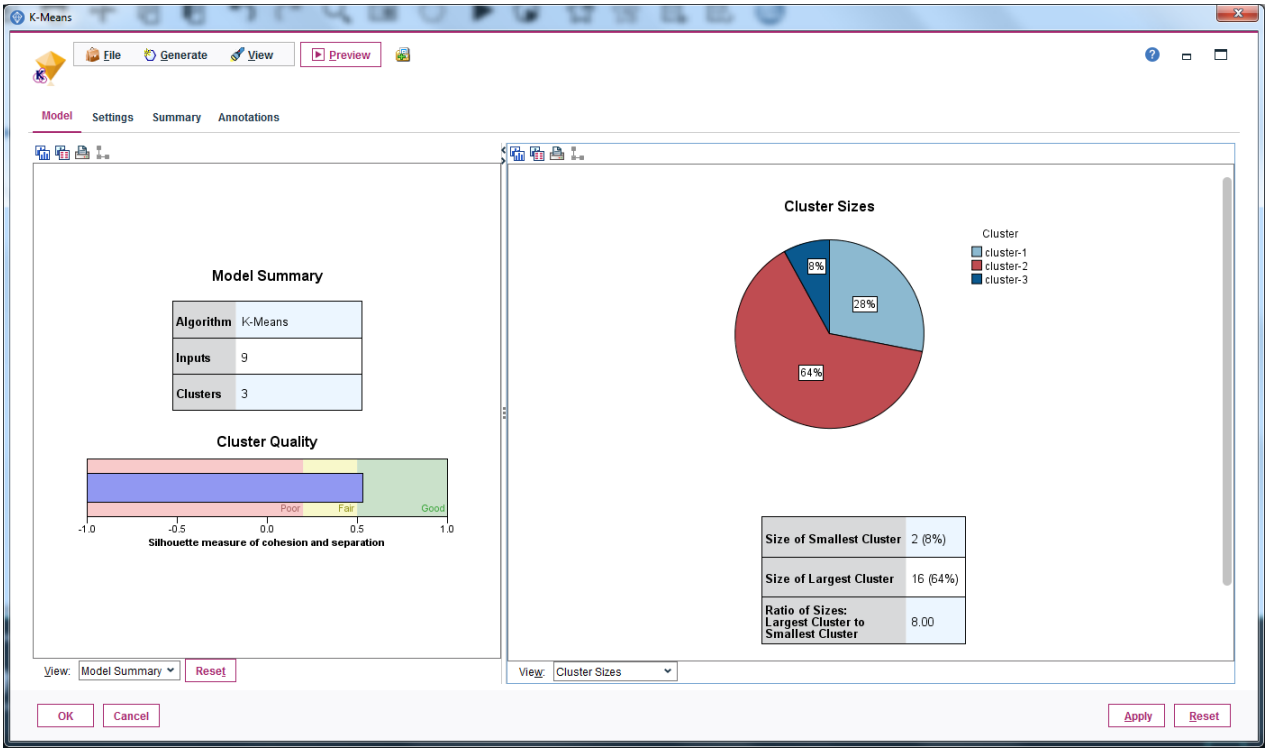
Duplim klikom na model klasterovanja može se videti rezultat klasterovanja i izvršiti detaljnija analiza izdvojenih klastera.

Na pogledu *Model Summary* vidi se da je silueta koeficijent 0,5, čime se smatra da je izvršeno dobro klasterovanje. Na pogledu *Predictor Importance* vidi se da je atribut *Nuts* najznačajniji za klasterovanje, zatim slede *Cereals, Starch, Milk, Eggs, Fish*, dok najmanji značaj imaju *WhiteMeat, Fr&Veg* i *RedMeat*(Slika 6).




**Slika 6:** Pogledi: *Model Summary* i *Predictor Importance*

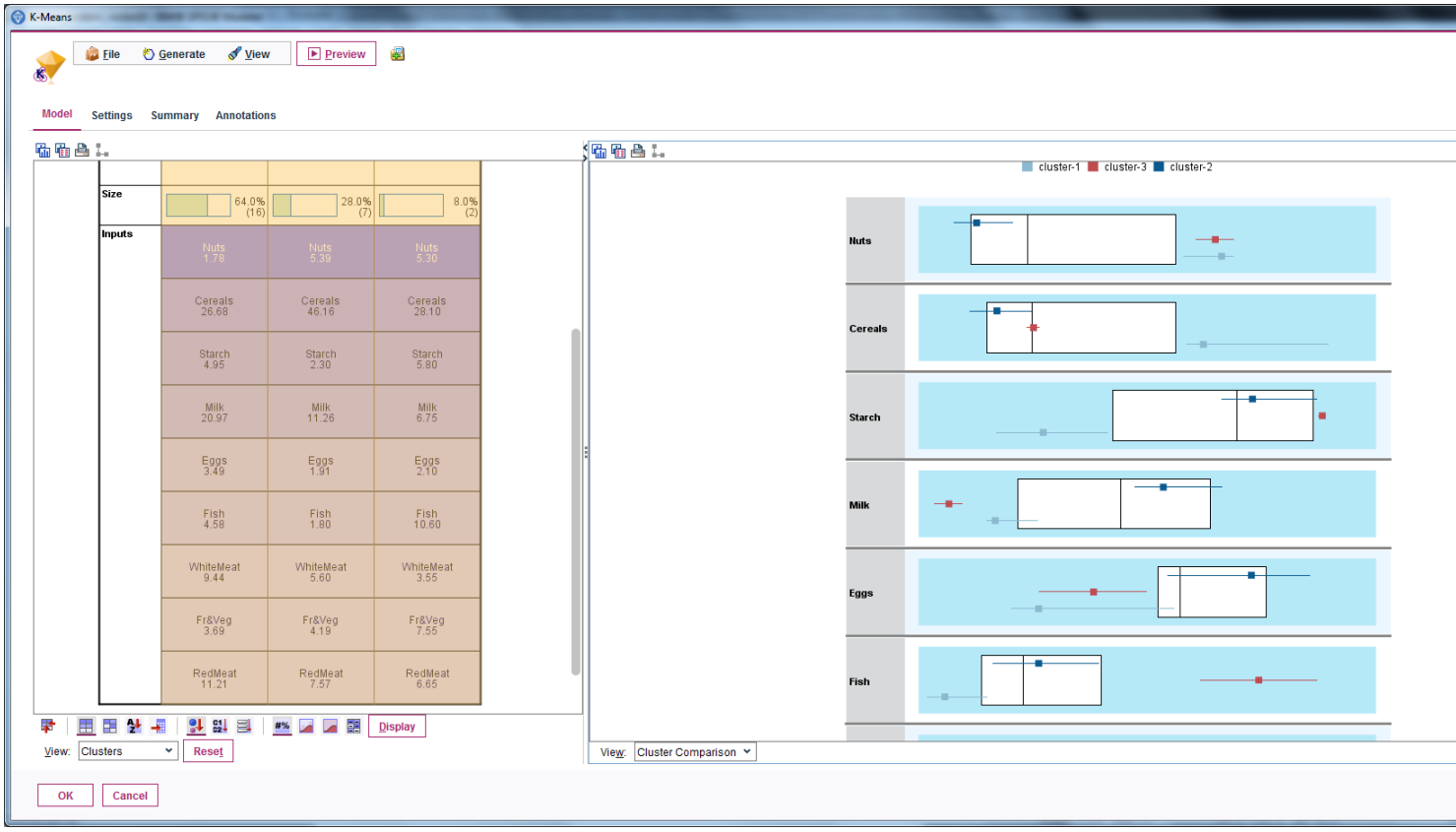
Na pogledu *Cluster Sizes* (Slika 7) vidi se da su izdvojeni jedan veliki klaster (sa 64% država, **cluster-2**), jedan srednji (sa 28% država, **cluster-1**) i jedan mali (sa 8% država, **cluster-3**).




**Slika 7:** Pogledi: *Clusters* i *Cluster Sizes*

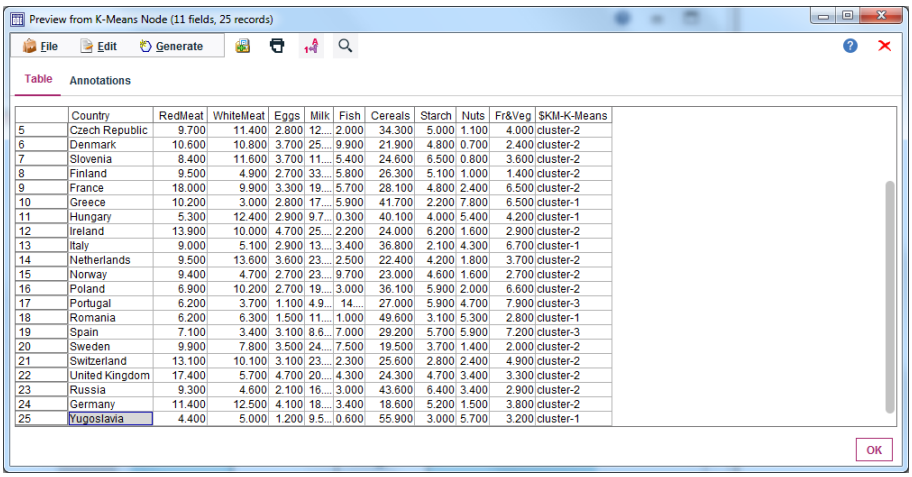
Preko pogleda *Clusters* i *Cluster Comparison* (Slika 8, za detaljniji uvid vrednosti atributa po klasterima pogledati radni tok) se može uočiti šta je specifično za svaki klaster:

- U ishrani stanovništva u državama najvećeg klastera, **cluster 2**, više je zastupljeno belo i crveno meso, mleko i jaja, dok su koštunjavo voće i žitarice manje zastupljeni u odnosu na države u ostalim klasterima.
- U ishrani stanovništva u državama srednjeg klastera, **cluster 1**, značajno su zastupljenije žitarice (srednja vrednost atributa *Cereals* za klaster 1 je 46,16, dok je za 2. i 3. klaster redom 26,68 i 28,1), dok su skrob i riba manje zastupljeni u odnosu na države u ostalim klasterima.
- U ishrani stanovništva u državama malog klastera, **cluster 3**, više je zastupljen skrob, kao i riba, voće i povrće, dok su mleko i belo meso manje zastupljeni u odnosu na države u ostalim klasterima.




**Slika 8:** Pogledi *Clusters* i *Cluster Comparison*

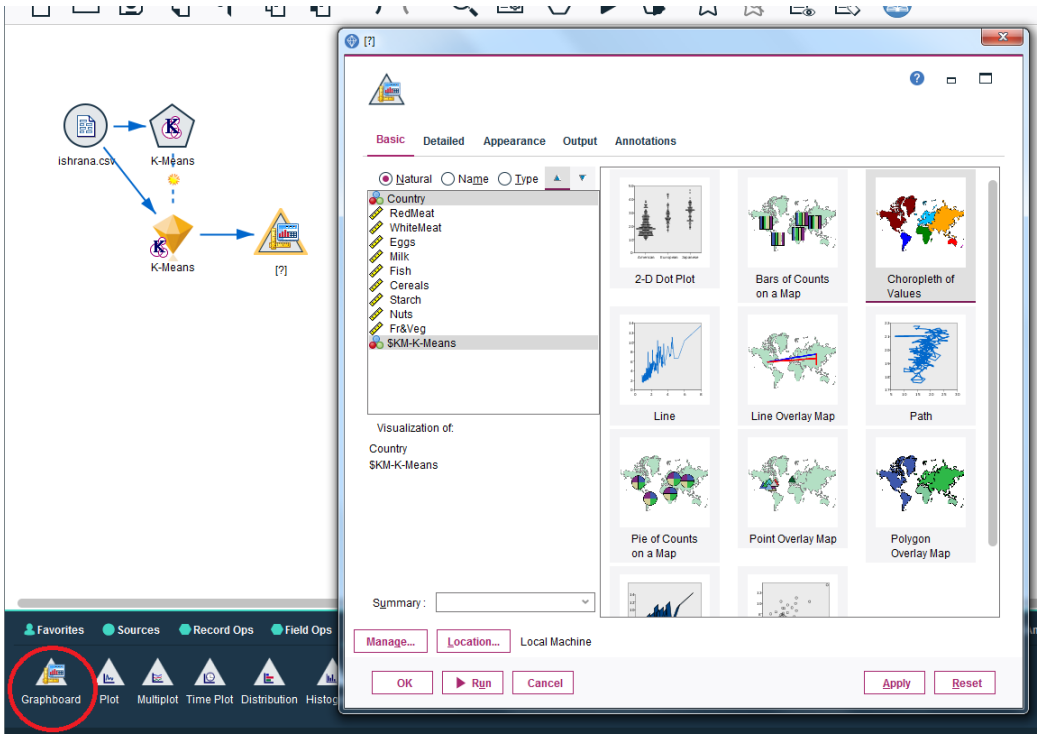
U modelu, klikom na dugme *Preview* može se videti za svaku instancu, tj. državu, kom klasteru je dodeljena (Slika 9).



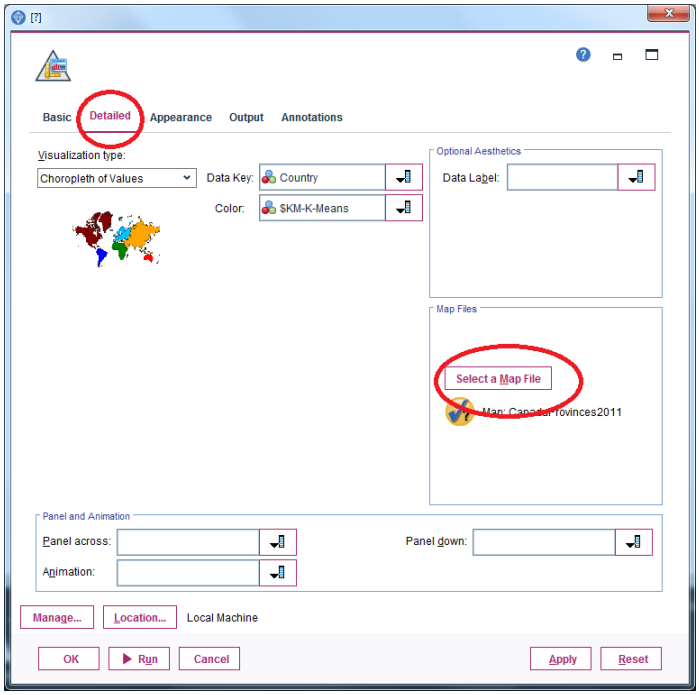

| | Country | RedMeat | WhiteMeat | Eggs | Milk | Fish | Cereals | Starch | Nuts | Fr&Veg | $KM-K-Means |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Czech Republic | 9.700 | 11.400 | 2.800 | 12... | 2.000 | 34.300 | 5.000 | 1.100 | 4.000 | cluster-2 |
| 6 | Denmark | 10.600 | 10.800 | 3.700 | 25... | 9.900 | 21.900 | 4.800 | 0.700 | 2.400 | cluster-2 |
| 7 | Slovenia | 8.400 | 11.600 | 3.700 | 11... | 5.400 | 24.600 | 6.500 | 0.800 | 3.600 | cluster-2 |
| 8 | Finland | 9.500 | 4.900 | 2.700 | 33... | 5.800 | 26.300 | 5.100 | 1.000 | 1.400 | cluster-2 |
| 9 | France | 18.000 | 9.900 | 3.300 | 19... | 5.700 | 28.100 | 4.800 | 2.400 | 6.500 | cluster-2 |
| 10 | Greece | 10.200 | 3.000 | 2.800 | 17... | 5.900 | 41.700 | 2.200 | 7.800 | 6.500 | cluster-1 |
| 11 | Hungary | 5.300 | 12.400 | 2.900 | 9.7... | 0.300 | 40.100 | 4.000 | 5.400 | 4.200 | cluster-1 |
| 12 | Ireland | 13.900 | 10.000 | 4.700 | 25... | 2.200 | 24.000 | 6.200 | 1.600 | 2.900 | cluster-2 |
| 13 | Italy | 9.000 | 5.100 | 2.900 | 13... | 3.400 | 36.800 | 2.100 | 4.300 | 6.700 | cluster-1 |
| 14 | Netherlands | 9.500 | 13.600 | 3.600 | 23... | 2.500 | 22.400 | 4.200 | 1.800 | 3.700 | cluster-2 |
| 15 | Norway | 9.400 | 4.700 | 2.700 | 23... | 9.700 | 23.000 | 4.600 | 1.600 | 2.700 | cluster-2 |
| 16 | Poland | 6.900 | 10.200 | 2.700 | 19... | 3.000 | 36.100 | 5.900 | 2.000 | 6.600 | cluster-2 |
| 17 | Portugal | 6.200 | 3.700 | 1.100 | 4.9... | 14... | 27.000 | 5.900 | 4.700 | 7.900 | cluster-3 |
| 18 | Romania | 6.200 | 6.300 | 1.500 | 11... | 1.000 | 49.600 | 3.100 | 5.300 | 2.800 | cluster-1 |
| 19 | Spain | 7.100 | 3.400 | 3.100 | 8.6... | 7.000 | 29.200 | 5.700 | 5.900 | 7.200 | cluster-3 |
| 20 | Sweden | 9.900 | 7.800 | 3.500 | 24... | 7.500 | 19.500 | 3.700 | 1.400 | 2.000 | cluster-2 |
| 21 | Switzerland | 13.100 | 10.100 | 3.100 | 23... | 2.300 | 25.600 | 2.800 | 2.400 | 4.900 | cluster-2 |
| 22 | United Kingdom | 17.400 | 5.700 | 4.700 | 20... | 4.300 | 24.300 | 4.700 | 3.400 | 3.300 | cluster-2 |
| 23 | Russia | 9.300 | 4.600 | 2.100 | 16... | 3.000 | 43.600 | 6.400 | 3.400 | 2.900 | cluster-2 |
| 24 | Germany | 11.400 | 12.500 | 4.100 | 18... | 3.400 | 18.600 | 5.200 | 1.500 | 3.800 | cluster-2 |
| 25 | Yugoslavia | 4.400 | 5.000 | 1.200 | 9.5... | 0.600 | 55.900 | 3.000 | 5.700 | 3.200 | cluster-1 |

**Slika 9:** Prikaz atributa koji je dodao model na originalan skup atributa

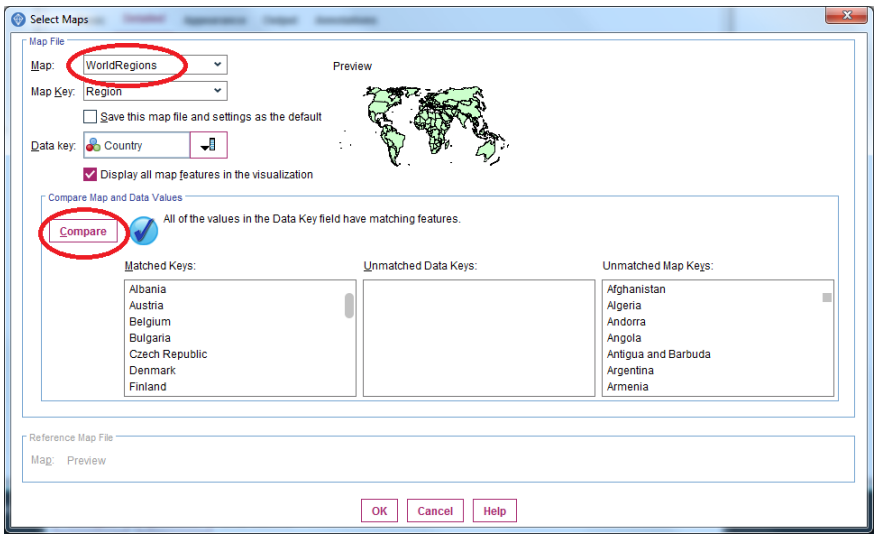
Pomoću horopleta (tematska mapa koja prikazuju neke informacije po jedinicama površine, bojenjem ili šrafiranjem) može se i vizuelno prikazati rezultat klasterovanja. Model klasterovanja je potrebno povezati sa čvorom za pravljenje grafika *Graphboard*. Odabirom atributa ime države (*Country*) i atributa koji sadrži podatak kom klasteru pripada instanca (*$KM-K-means*) na levoj strani čvora *Graphboard*, na desnoj strani čvora se prikazuju grafici pogodni za vizuelizaciju, među kojima je i horoplet (*Choropleth of Values*) (Slika 10). U odeljku *Detailed* proveriti da li su dobro postavljeni parametri. Za vrednost parametra *Data Key* postaviti atribut koji sadrži podatak o jedinici površine, u ovom primeru to je atribut *Country*. Za vrednost parametra *Color* postaviti atribut *$KM-K-means*, čime se svakom klasteru dodeljuje jedinstvena boja, a svaka država na mapi se boji bojom koja je dodeljena klasteru kome pripada (Slika 11). Klikom na dugme *Select a Map File* otvara se dijalog u kome je potrebno izabrati željenu mapu. Za ovaj zadatak, to je mapa sveta, te se za parametar *Map* postavlja vrednost *WorldRegions*. Da bi se uparili podaci iz postavljenog ključa (u ovom primeru atribut *Country*) sa jedinicama na mapi, kliknuti na dugme *Compare*. U listi *Matched Keys* prikazuju se imena država iz skupa koje su uparene sa ključem neke jedinice sa mape (u ovom primeru će biti sve države iz skupa). U listi *Unmatched Keys* prikazuju se imena država iz skupa koje nisu uparene sa ključem neke jedinice sa mape (u ovom primeru nema takvih država). U listi *Unmatched Map Keys* prikazuju se ključevi jedinica sa mape koji nisu upareni sa nekom instancom iz skupa (pošto skup sadrži podatke za 26 država, u ovoj listi će biti veliki broj takvih država) (Slika 12).




**Slika 10:** Izbor horopleta za prikaz rezultata klasterovanja



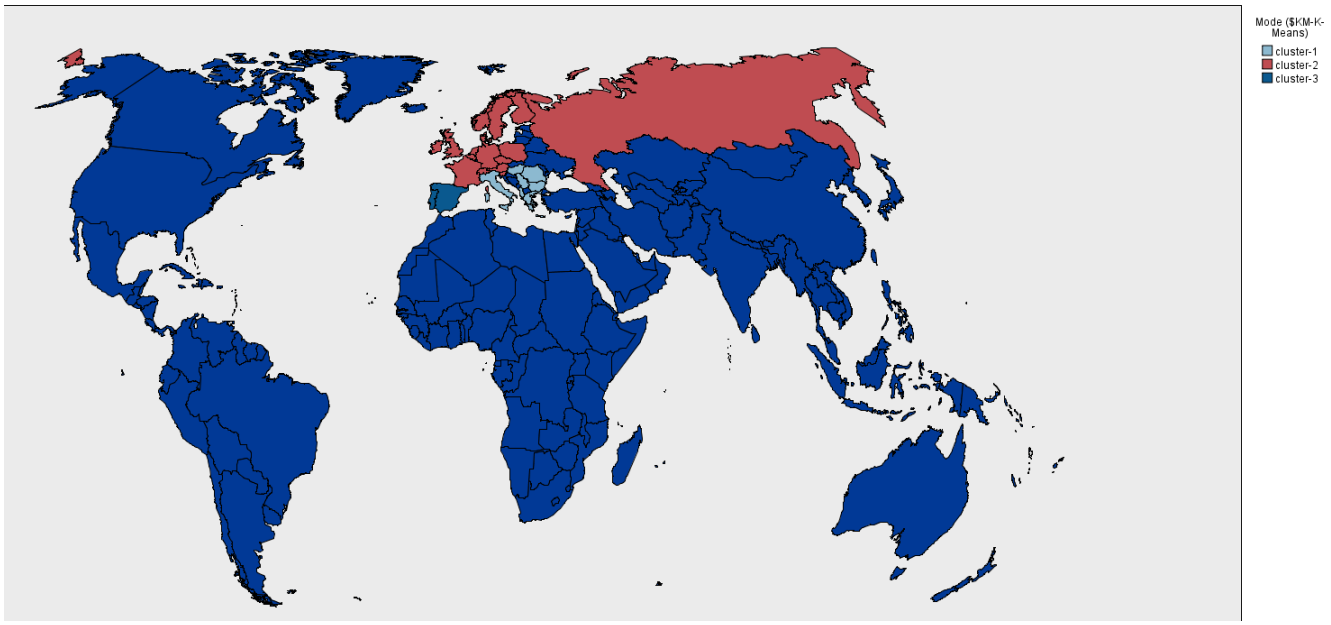

**Slika 11:** Dodela jedinstvene boje svakom klasteru zbog prikaza rezultata klasterovanja




**Slika 12:** Odabir mape za prikaz podataka

Na slici 13 je horoplet koji prikazuje rezultat klasterovanja. Najtamnijom plavom bojom su prikazane države za koje ne postoje podaci u skupu. Na osnovu mape može se videti koje države pripadaju kom klasteru:

- **cluster 1**: Italija, Rumunija, Grčka, Mađarska, Srbija (u skupu je promenjena na Jugoslavija zbog mape), Albanija i Bugarska
- **cluster 3**: Španija i Portugal
- **cluster 2**: ostale zemlje iz skupa




**Slika 13:** Prikaz rezultata klasterovanja pomoću mape sveta

Za potrebe tekstualnog izveštaja klasterovanja, primenom čvorova *Sort* i *Table* mogu se urediti instance skupa prema klasteru kome pripadaju (videti radni tok).

Promenom broja željenih klastera na vrednost u intervalu od 4 do 9 silueta koeficijent klasterovanja se ne menja, tj. ostaje 0,5. Ako se postavi broj željenih klastera na 10, silueta koeficijent se povećava na 0,6. Međutim, za skup od 26 instanci, 10 je veliki broj klastera, posebno što tri dobijena klastera sadrže po jednu državu.