Tehnike i metode analize podataka

**Klasterizacija podataka o gojaznosti korišćenjem KMeans i Agglomerative clustering algoritama**

Danilo Milošević 1732

**Opis podataka i problema**

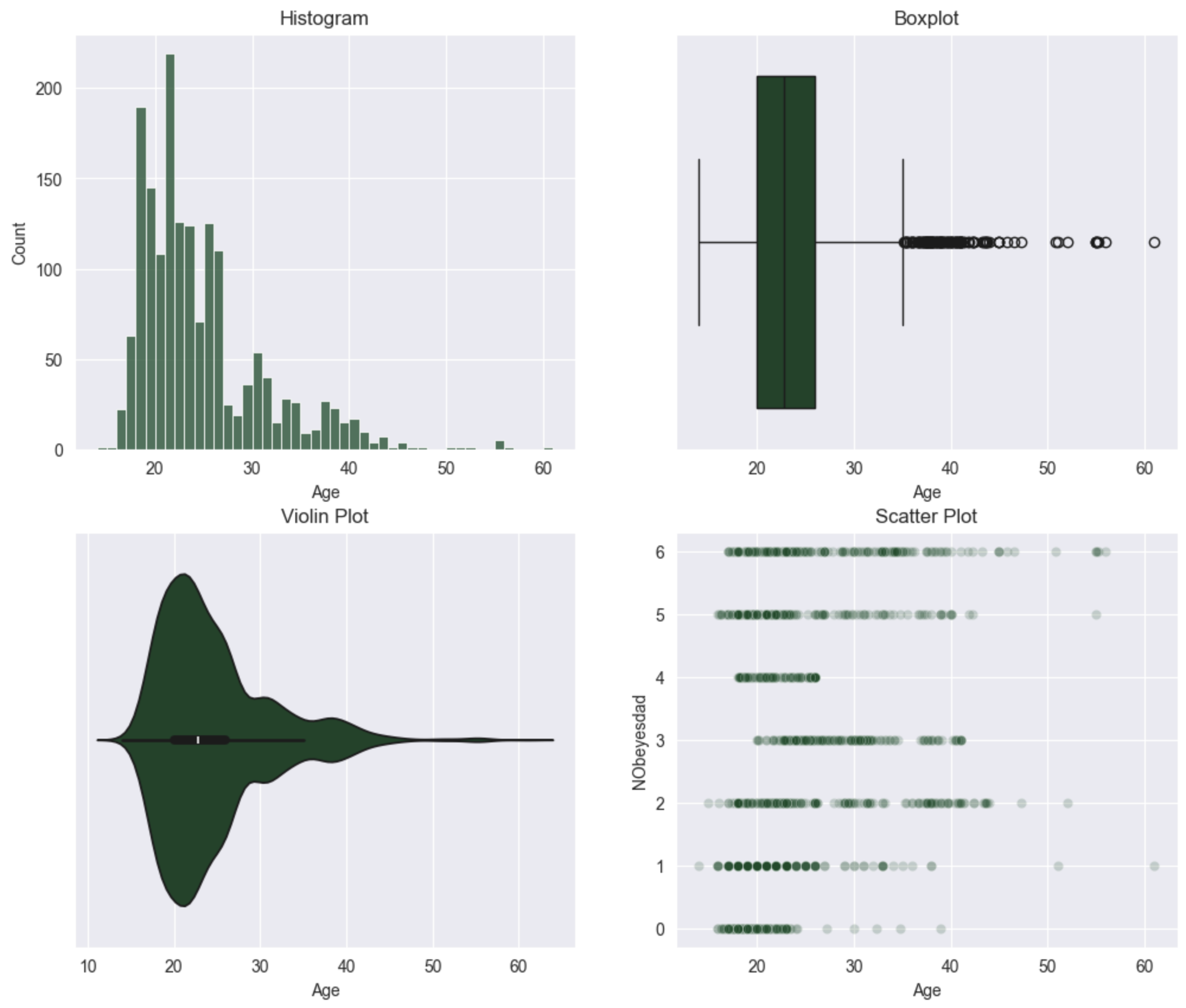
Vrši se klasterizacija na osnovu skupa podataka koji sadrži zdravstvene podatke o ljudima kao što su visina, pol i kilaža sa ciljem podele ljudi u grupe, analizom datih grupa kao i predviđanja nivoa gojaznosti na osnovu datih podataka. Atributi su sledeći:

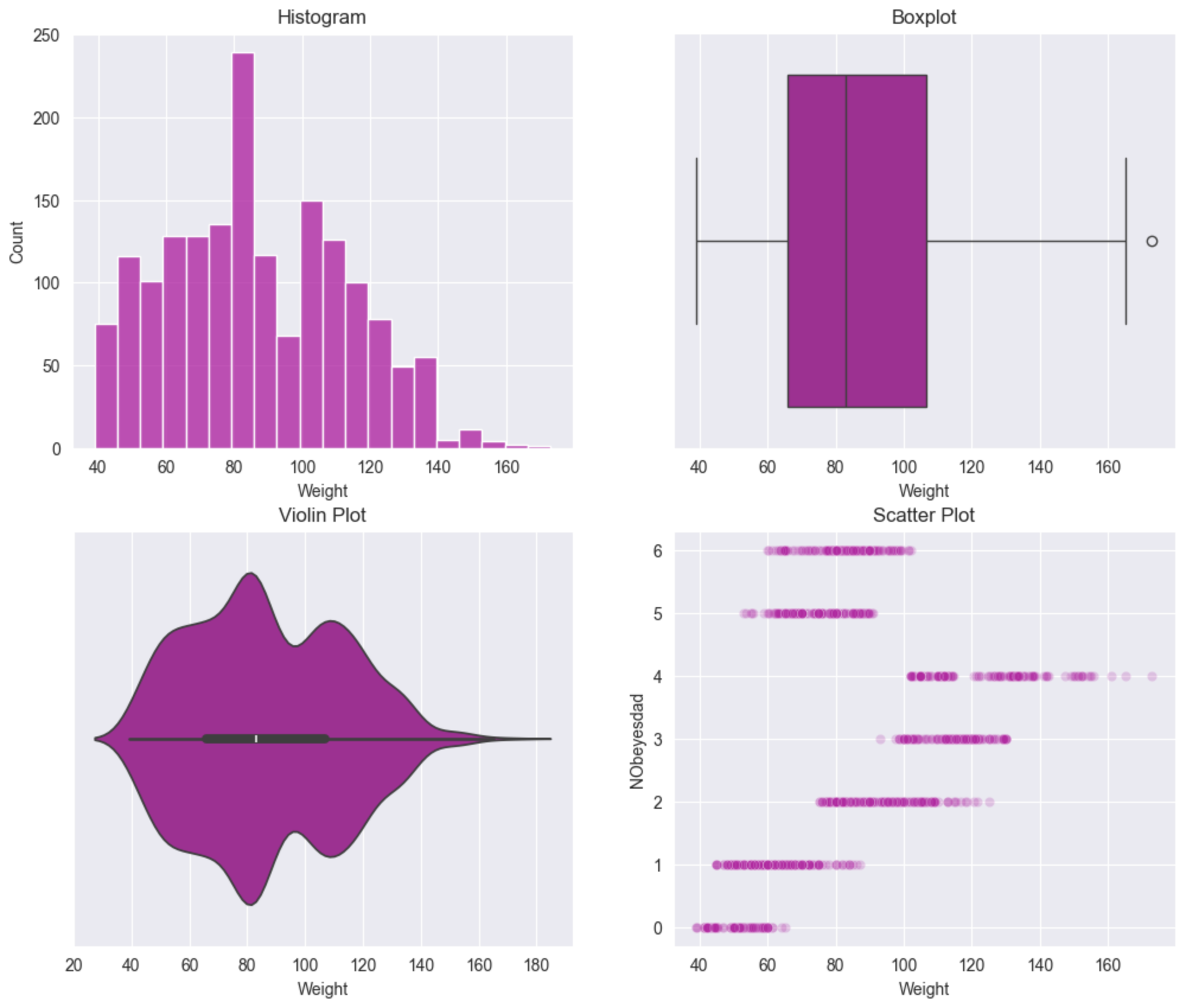
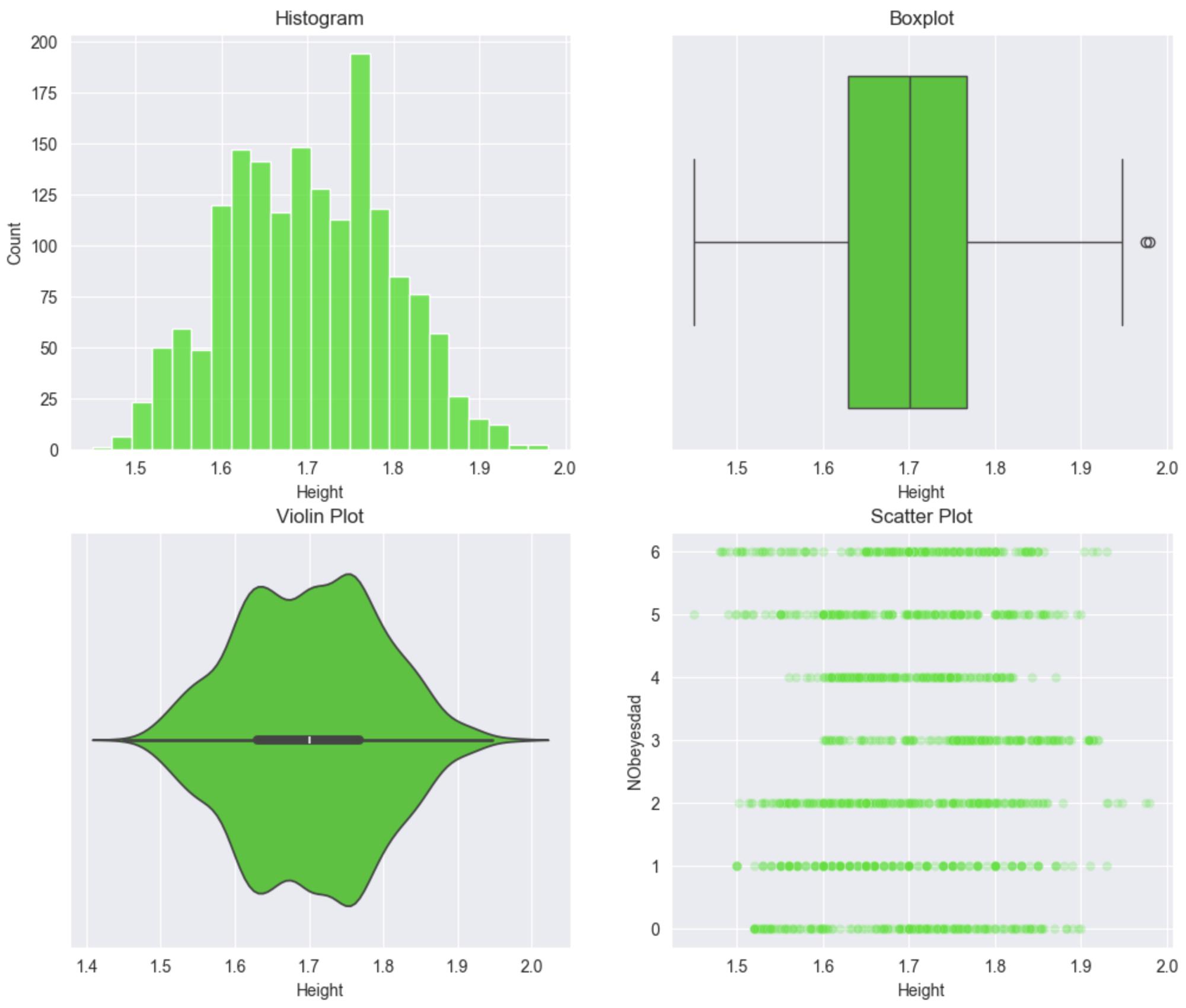
* Gender - pol
* Age - starost
* Height - visina
* Weight - kilaža
* family\_history\_with\_overweight - da li su članovi porodice gojazni (yes-no)
* FAVC - čest unos visoko kalorične hrane
* FCVC - učestalost unosa povrća
* NCP - broj glavnih jela
* CAEC - unos hrane između obroka - 4 jedinstvene vrednosti
* SMOKE - pušač (da-ne)
* CH2O - unos vode u toku dana
* SCC - da li vode računa o unosu kalorija (da-ne)
* FAF - učestalost fizičke aktivnosti
* TUE - vreme korišćenja tehnologije
* CALC - unos alkohola - 4 jedinstvene vrednosti
* MTRANS - korišćenje prevoznih sredstava - 5 jedinstvene vrednosti
* NObeyesdad - Nivo gojaznosti - 7 jedinstvenih vrednosti

Skup sadrži 2111 podataka i 18 promenljivih. Za trening i test koristimo podelu 80:20. U skupu nema nedostajućih, odnosno null vrednosti. Neki od atributa su tipa object i moraju se enkodirati - Gender, family\_history\_with\_overweight, FAVC, CAEC, SMOKE, SCC, CALC, MTRANS i NObeysdad.

Gender, family\_history\_with\_overweight, FAVC, SMOKE i SCC kodiramo sa 0 i 1 - true ili false, dok preostale kodiramo OneHot enkodiranjem, izuzev NObeysdad koji će biti kodiran label encoder-om.

Kako ima previše atributa, ispod je prikazana vizuelizacija samo prvih par - Gender, Age, Height i Weight. Na osnovu grafikona možemo da zaključimo da skup podataka čini relativno jednak broj muškaraca i žena, da su mladi ljudi dosta zastupljeniji od starijih, da visina varira između 1.5m i 1.9m pri čemu su viši od toga retki. Pored toga kilaža varira od 40kg do 160kg, dok je najzastupljenija kilaža oko 80kg.





**KMeans**

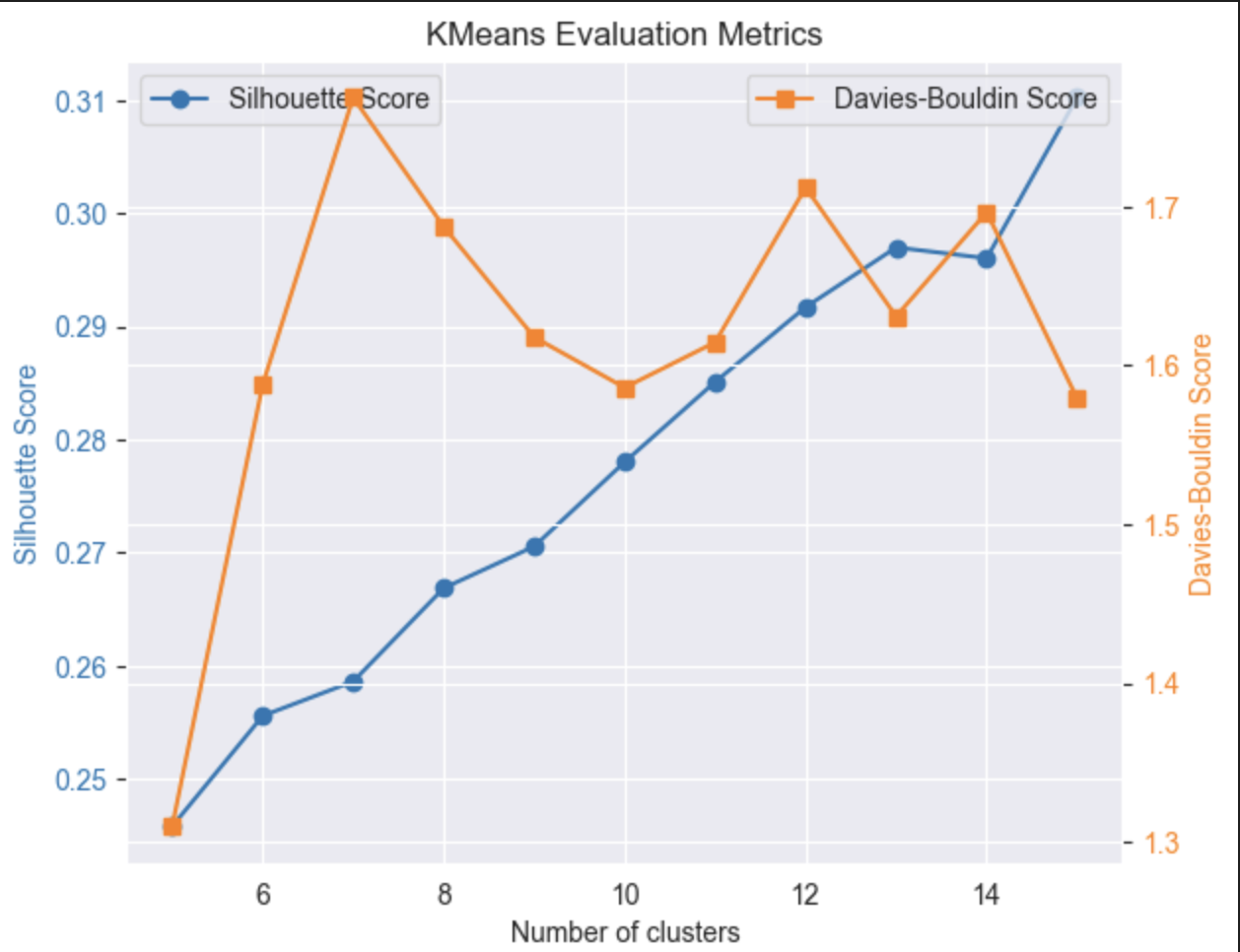
Klasterizaciju vršimo prvo algoritmom KMeans, koji ćemo trenirati korišćenjem viša kombinacija transformacija nad ulaznim podacima, odrediti silhoutte score kao i Davis-Bouldin score kako bi izabrali najbolji broj cluster-a. Zatim vršimo i vizuelizaciju klasterizacije pomoću tSNE algoritma. Pored toga ćemo prikazati i raspodelu svih atributa po clusterima kako bi analizirali rezultate. Kako ima puno dijagrama prikazaćemo samo score-ove na osnovu kojih je izbrana kombinacija transformacija i broj clustera. Cilj je minimizovati Davies-Bouldin score i maksimizovati silhoutte score.

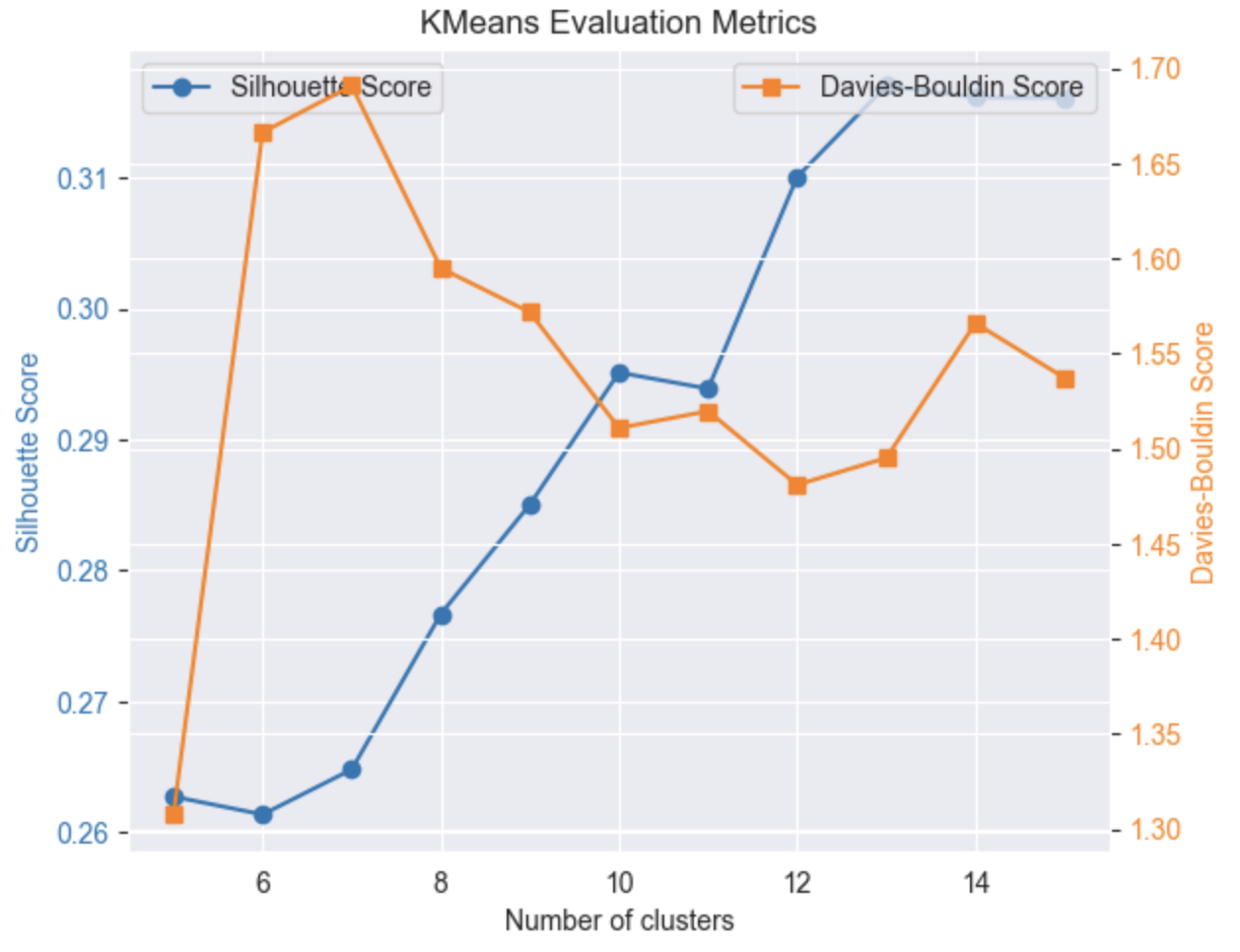
Prvi slučaj je primena kmeans na skup podataka bez transformacija.



Najbolji rezultat je za 6 clustera, međutim biramo 7 jer se poklapa sa brojem grupa gojaznosti a rezultati su približno jednaki.

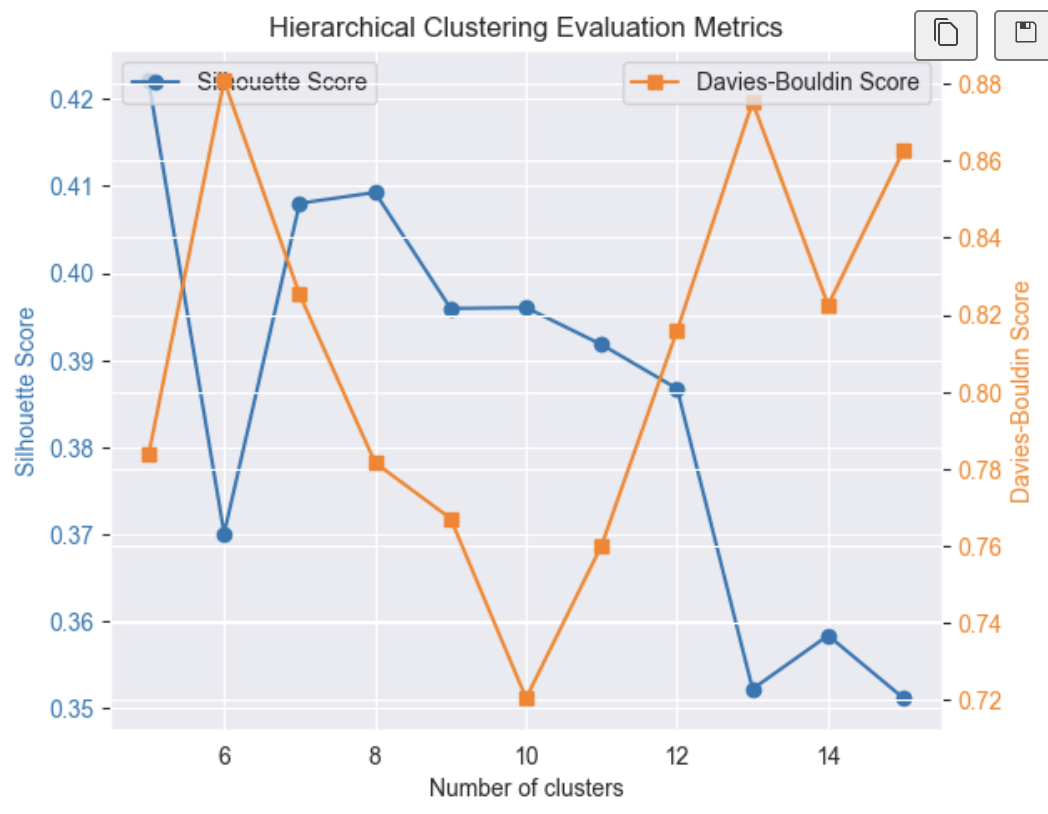
Sledeći slučajevi daju dosta gore rezultate.





*Sleva udesno - sa skaliranjem, sa skaliranjem i normalizacijom, sa skaliranjem, normalizacijom i PCA, skaliranjem i PCA*

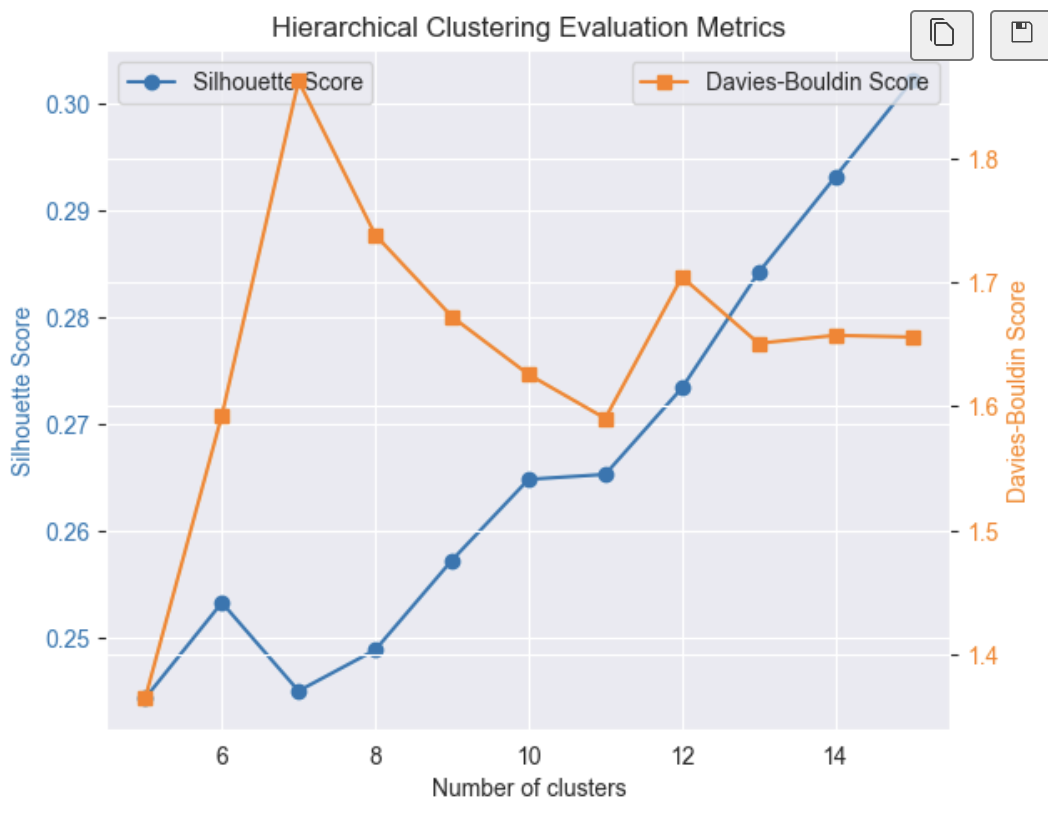
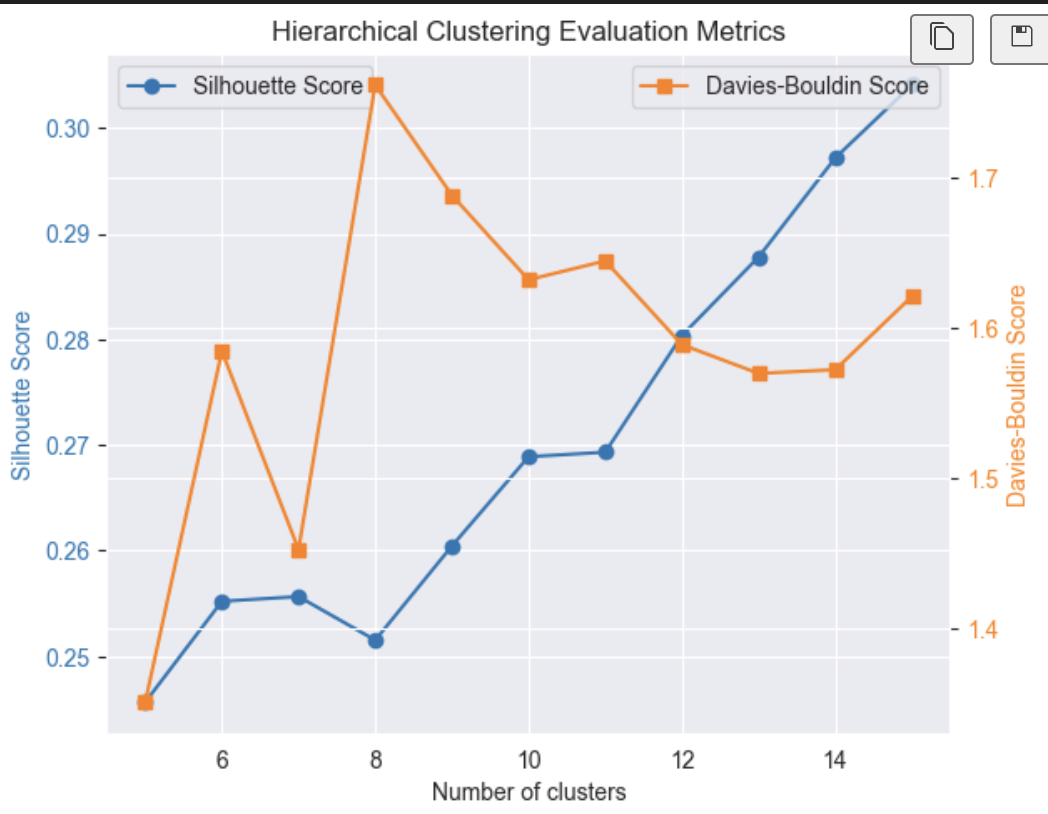
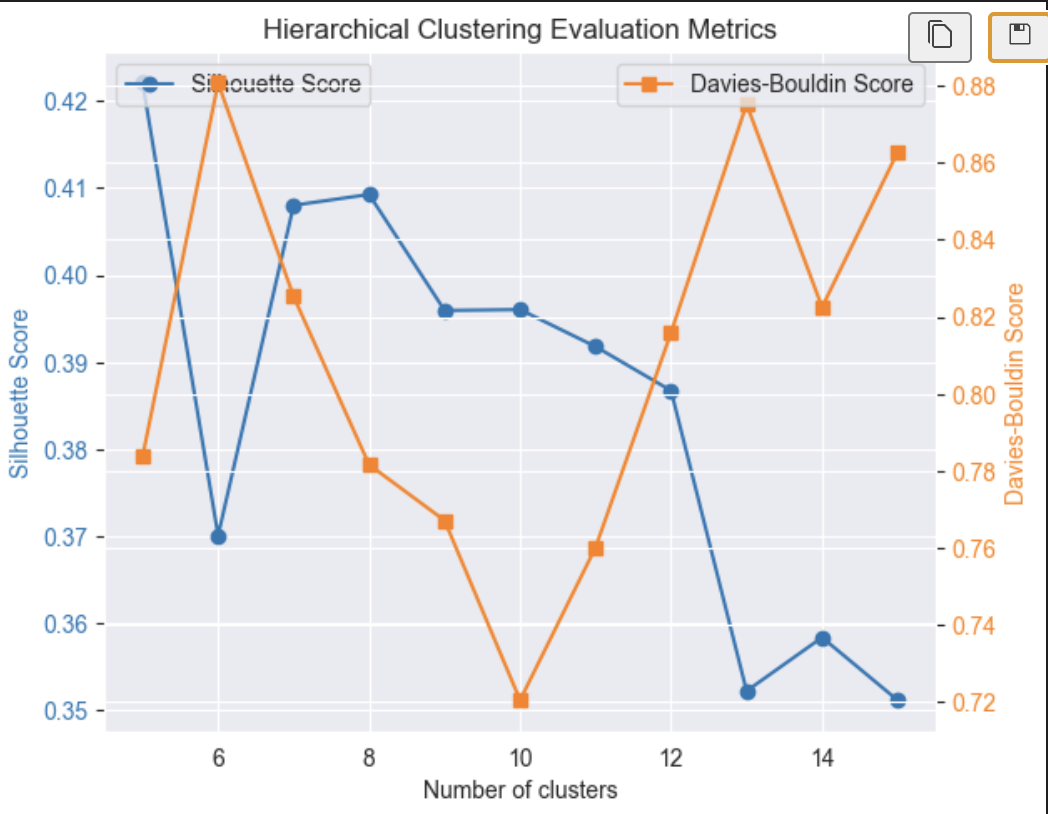
Najbolje rezultate je dao kmeans u kombinaciji sa PCA.

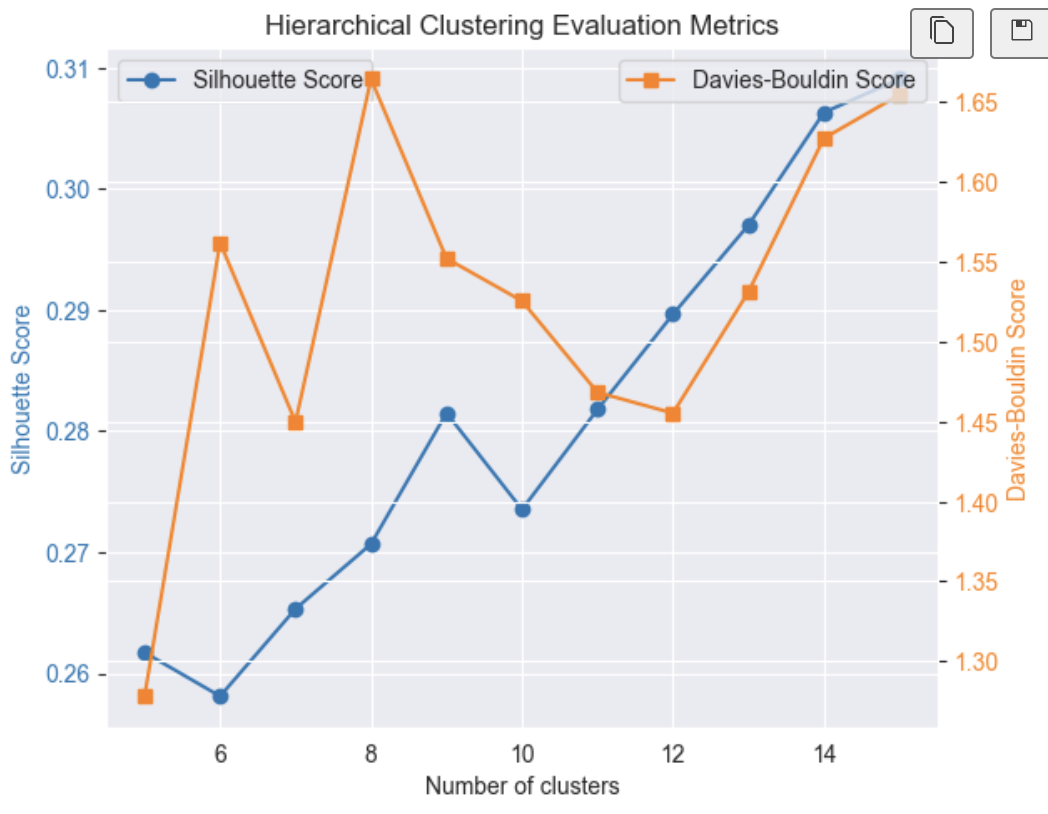
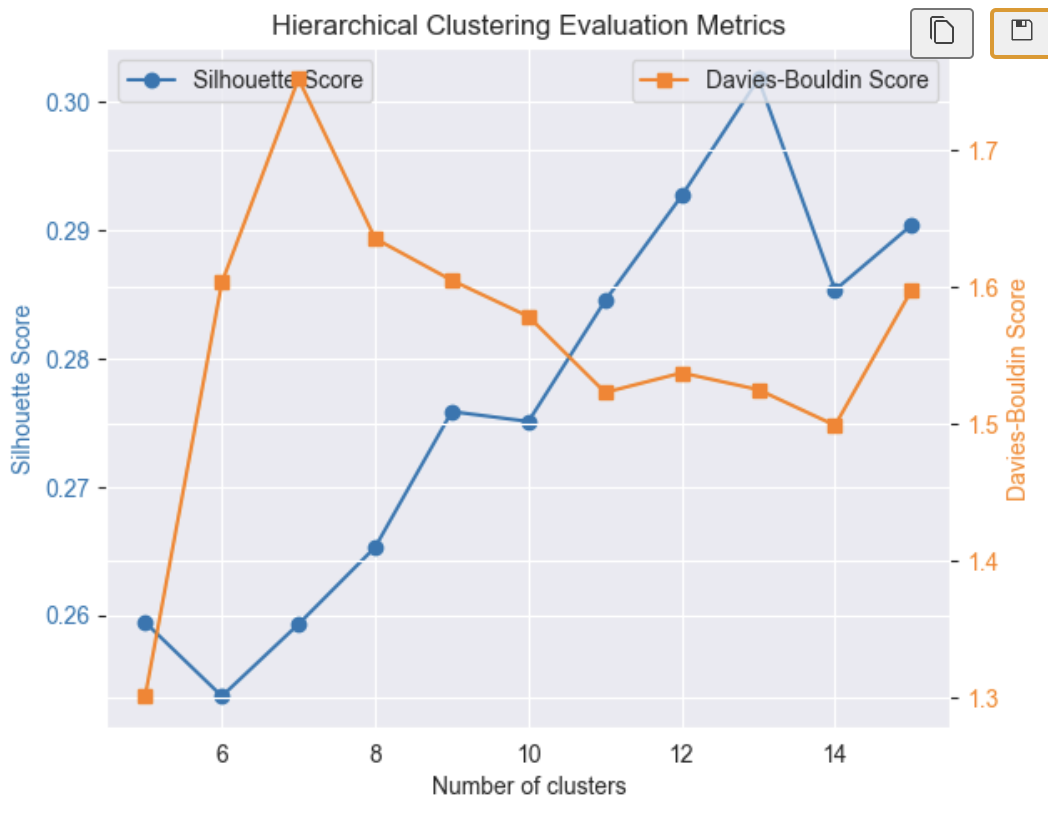


**Agglomerative clustering**

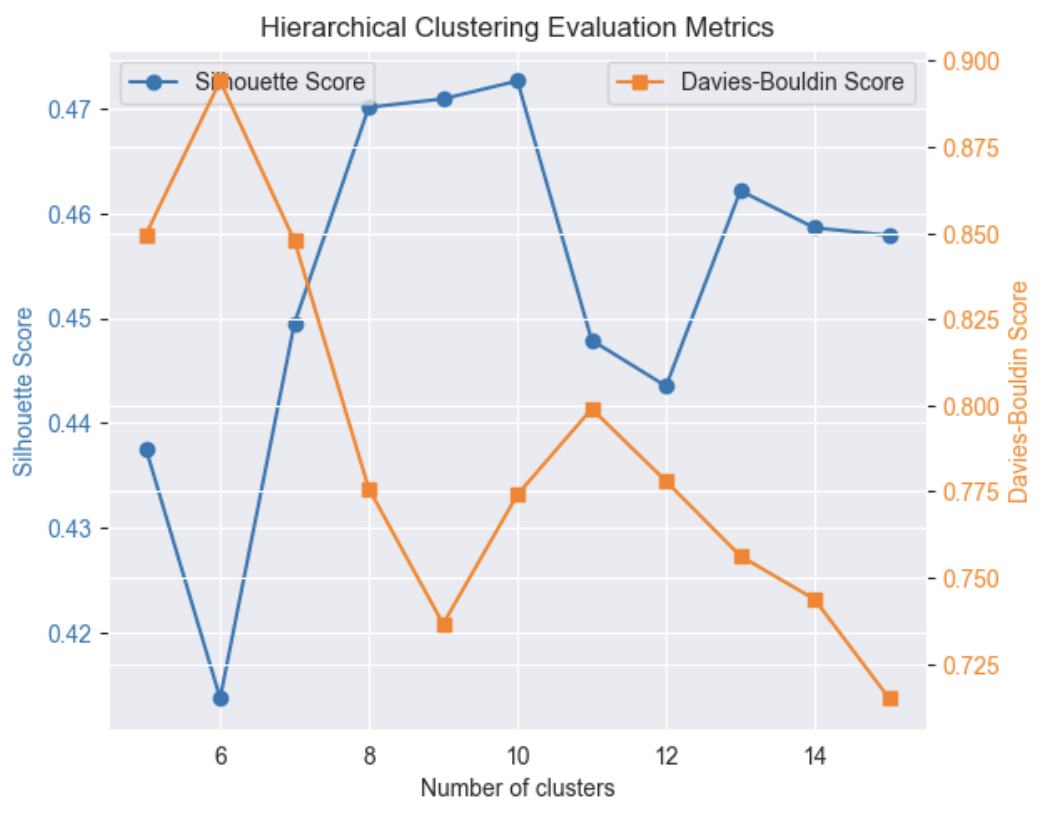
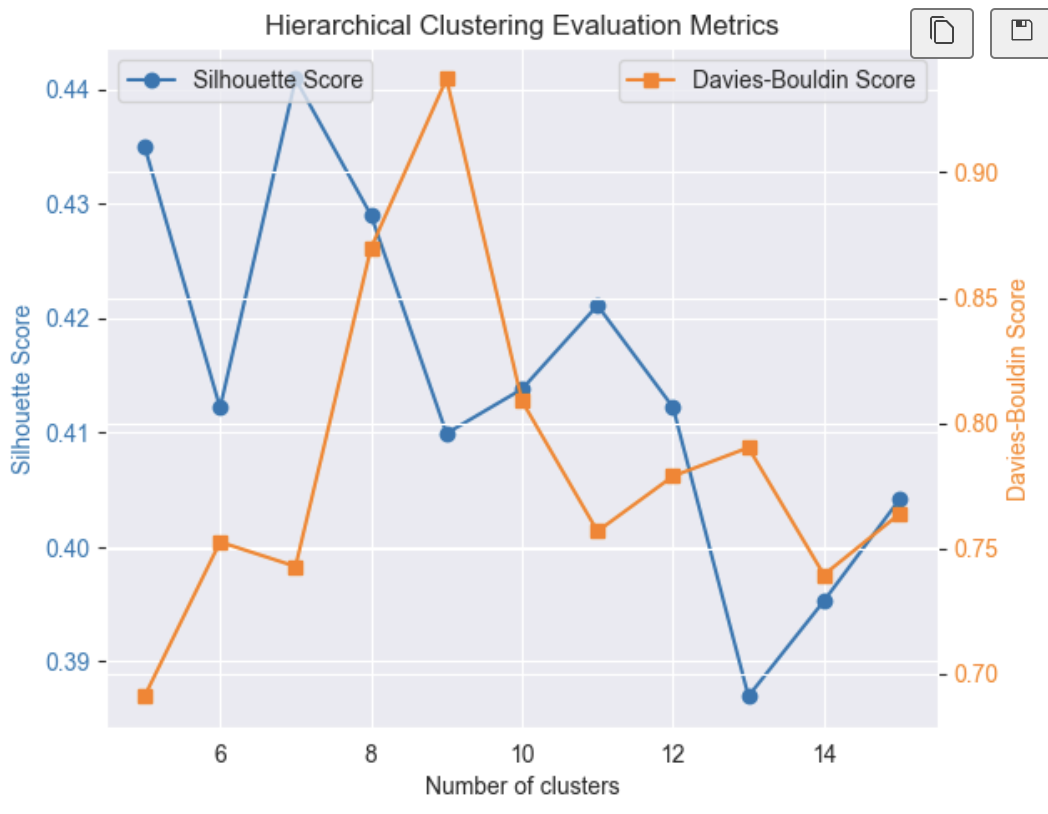
Klasterizaciju vršimo zatim algoritmom Agglomerative clustering, koji ćemo trenirati korišćenjem viša kombinacija transformacija nad ulaznim podacima, odrediti silhoutte score kao i Davis-Bouldin score kako bi izabrali najbolji broj cluster-a. Zatim vršimo i vizuelizaciju klasterizacije pomoću tSNE algoritma. Pored toga ćemo prikazati i raspodelu svih atributa po clusterima kako bi analizirali rezultate. Kako ima puno dijagrama prikazaćemo samo score-ove na osnovu kojih je izbrana kombinacija transformacija i broj clustera. Cilj je minimizovati Davies-Bouldin score i maksimizovati silhoutte score.

Agglomerative clustering je dao gore rezultate od KMeans-a osim u slučaju kada je primenjena redukcija dimenzionalosti - PCA kao i Isomap.





*Sleva udesno - bez transformacija, sa skaliranjem, sa skaliranjem i normalizacijom, sa skaliranjem, normalizacijom i PCA, skaliranjem i PCA*



PCA i Isomap zajedno sa Agglomerative clustering su dali najbolje rezultate. Rezultati su vrlo slični, dok su performanse PCA dosta bolje od Isomap, zbog čega je kao najbolji model izabran PCA u kombinaciji sa Agglomerative clustering.

**Analiza clustera**

Detaljna analiza i grafikoni raspodele atributa po clusterima je dat u Jupyter svesci, dok se ovde nalaze samo krajnji opisi clustera. Korišćen je agglomerative clustering algoritam u kombinaciji sa PCA, za 7 clustera.

* Cluster 0
  + Cluster čine mladi niži ljudi, kilaže najčešće 70-85kg, zbog čega je nivoa gojaznosti visok. Ostale karakteristike su prosečne (unos alkohola, tehnologije, transporta, unos vode i drugih atributa).
* Cluster 1
  + Cluster čine starije osobe, prosečno visoke sa velikom kilažom. Većina osoba se hrane kaloričnom hranom i imaju problema sa kilažom u porodici. Nisu pušači, često uopšte ne piju alkohol, ali se slabo bave fizičkom aktivnošću i uglavnom se kreću automobilom. Gojaznost je zastupljena u ovom clusteru.
* Cluster 2
  + Cluster uglavnom čine mlade odrasle osobe, pretežno muškarci, nadprosečne visine sa većom kilažom. Sve osobe se hrane kaloričnom hranom i imaju problema sa kilažom u porodici. Pored toga retko se bave fizičkom aktivnošću i kreću se najčešće javnim prevozom ili automobilom. Gojaznost je zastupljena u ovom clusteru, i to varira od blago povišene do gojaznosti prvog stepena.
* Cluster 3
  + Cluster uglavnom čine mlade ženske osobe, niske-prosečne visine sa malom kilažom. Sve osobe se hrane kaloričnom hranom, često unose povrće i često nemaju problema sa kilažom u porodici. Pored toga retko se bave fizičkom aktivnošću i kreću se najčešće javnim transportom. U okviru clustera nije prisutna gojaznost.
* Cluster 4
  + Cluster uglavnom čine mladi ljudi, uglavnom 25-godišnjaci, niske-prosečne visine sa velikom kilažom. Većina osoba se hrani kaloričnom hranom i imaju problema sa kilažom u porodici. Pored toga vrlo retko se bave fizičkom aktivnošću i obično koriste javni prevoz. U okviru clustera je prisutna umerena kao i ekstremna gojaznost.
* Cluster 5
  + Cluster uglavnom čine mladi i srednjovečni muškarci, niske-prosečne visine sa normalnom kilažom. Većina osoba se hrani kaloričnom hranom ali i povrćem i imaju problema sa kilažom u porodici. Pored toga retko se bave fizičkom aktivnošću (bolje od drugih clustera), često koriste tehnologiju. U okviru clustera je prisutna blago normalna kilaža kao i visoka gojaznost.
* Cluster 6
  + Cluster uglavnom čine mladi ljudi, malo češće žene, prosečne visine sa blago višom kilažom. Većina osoba se hrani kaloričnom hranom, donekle i povrćem i imaju problema sa kilažom u porodici. Pored toga retko-ponekad se bave fizičkom aktivnošću, vrlo često koriste tehnologiju. U okviru clustera je prisutna blaga kao i ekstremna gojaznost.