O datasetu:

Dataset se odnosi na klijente distributer ana veliko. Uključuje godišnju potrošnju u monetarnim jedinicama (m. u.) na različite kategorije proizvoda. Sadrži 440 instanci I 8 atributa:

FRESH: godišnja potrošnja (m. u.) na sveže proizvode (kontinualni atribut)

MILK: godišnja potrošnja (m. u.) na mlečne proizvode (kontinualni atribut)

GROCERY: godišnja potrošnja (m. u.) na namirnice (kontinualni atribut)

FROZEN: godišnja potrošnja (m. u.) na zamrznute proizvode (kontinualni atribut)

DETERGENTS\_PAPER: godišnja potrošnja (m. u.) na deterdžent i papirne proizvode (kontinualni atribut)

DELICATESSEN: godišnja potrošnja (m. u.) na delikates proizvode (kontinualni atribut)

REGION: Lisbon, Oporto ili Drugi (nominalni atribut)

CHANNEL: Horeca(Hotel/Restaurant/Cafe) ili Retail (nominalni atribut)

O radu:

Algoritam za klasterovanje koji je implementiran je KMeans i kodiran je u Python programskom jeziku u Jupyter Notebook-u. Izvršena je analiza nad dataset-om i utvrđeno je da nema nedostajućih vrednosti a ni duplikata. Enkodirane su vrednosti REGION i CHANNEL atributa uz pomoć One Hot Encoder-a, budući da su kategoričkog tipa da ne bi ometali algoritam klasterovanja koji primenjuje Euklidsku distancu u prostoru.

KMeans algoritam koji je implementiran u funkciji *kmeans(data, n\_clusters, no\_of\_itterations*) u početku, uz pomoć numpy funkcije *np.random.choice()* bira *, n\_clusters* inicijalnih centroida. Kasnije računa euklidske distance svih instanci dataseta od svih centroida i instance se dodeljuje onom centroidu kom je najbliža. Nakon toga se centroidi ponovo računaju kao srednja vrednost atributa tačaka koje mu pripadaju. Postupak se ponavalja dok ne istekne broj iteracija (*no\_of\_itterations*). Funkcija vraća labele za svaku od instanci koje predstavljaju idx klastera kome pripadaju.

Odmah posle toga, dodeljuje se datasetu kolona Clusters kojoj se dodaju prethodno pomenute labele. Na osnovu njih se vrši PCA analiza sa 3 PCA komponente, da bi se lakše predstavile instance u prostoru. Ukoliko se izvrši kmeans, jasno je da algoritam grupise instance po X osi na dijagramu. Na osnovu toga, izabran je broj klastera 6, i na dijagramu se vidi da algoritam obavlja solidan posao.

