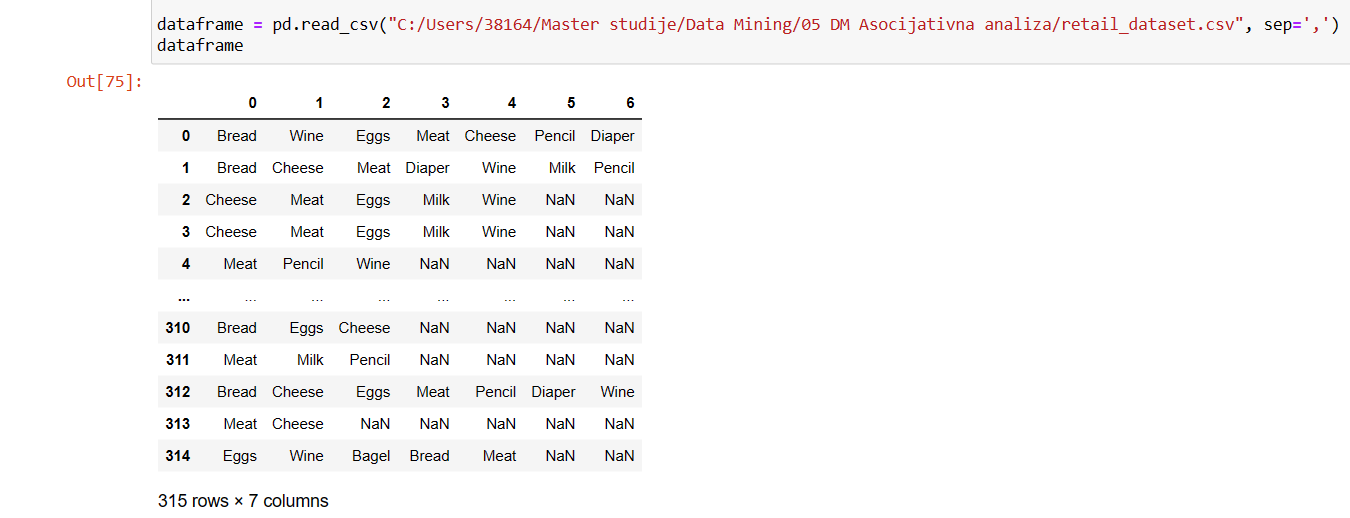
O datasetu:

Dataset predstavlja skup od 315 transakcija. Atributi su brojevi od 0 do 6 koji oznacavaju da transakcija moze imate maksimalno 7 stavki. Ukoliko u nekoj transakciji nema 7 stavki, u kolonama u Kojima nema stavki nalaze se NaN vrednosti. Stavki ima 9 i one su ‘Pencil’, ‘Eggs’, ‘Meat’, ‘Wine’, ‘Bagel’, ‘Bread’, ‘Cheese’, ‘Diaper’, ‘Milk’.



O radu:

Na samom pocetku, NaN vrednosti su zamenjene nulama. Nakon toga izvucene su sve stavke i smestene u listu *items.* Nakon toga je dataframe koji sadrzi dataset pretvoren u listu listi zbog laseg baratanja transakcijama u Apriori algoritmu.

Klasa je skoro savrseno balansirana:

**Implementacija Apriori algoritma:**

Prvi deo Apriori algoritma je generisanje frekventnog skupa stavki koje je postignuto implementacijom pomocne funkcije *generate\_frequent\_itemset(list\_of\_transactions, minsup)* koja kao argumente uzima listu transakcija i minimalnu potporu koju transakcija mora da predje kako bi bila smatrana frekventnom. Funkcija na pocetku od lista stavki izdvaja one koje su frekventne. Potom se na osnovu njih pune dve liste: lista frekventnih skupova stavki i lista nefrekventnih skupova stavki na osnovu koje mozemo unapred znati da li je neki skup stavki frekventan ili ne. Ukoliko je neki element nefrekventnog skupa stavki podskup skupa stavki koji razmatramo, za taj skup stavki mozemo odmah da kazemo da je nefrekventan i nema potrebe racunati potporu za njega i dodaje se u skup nefrekventnih stavki. Ovaj proces se zaustavlja obrade pravila kardinalnosti jednakoj broju frekventnih stavki kardinalnosti 1.

Drugi deo Apriori algoritma je generisanje pravila koje je postignuto funkcijom *generate\_rules(freq\_itemset, minsup, minconf)*. Funkcija poziva funkciju za generisanje frekventnog skupa stavki i na osnovu njih formira moguca pravila koja moraju imati pouzdanost vecu od *minconf.* Za svaki ferkventan skup stavki se generisu sve moguce desne strane asocijativnog pravila (*all\_possible\_y*). Za svaku od mogucih desnih strana jednog pravila se kreira i leva strana pravila (*x*). Na osnovu njih se racuna pouzdanost tog pravila, i ako je ono vece od minimalnog praga pouzadnosti, dodaje se u recnik konacnih pravila, a ako ne, dodaje se u listu pravila sa niskom pouzdanoscu (*low\_confidence\_rules*) na osnovu koje je moguce unapred odseci pravila sa niskom pouzdanosu (ukoliko je neki od elemenata low\_confidence\_rules liste podskup od desne strane pravila koje se razmatra).

Pomenuta odsecanja znacajno mogu umanjiti vreme izracunavanja kada je skup stavki veliki. Testirane su funkcije sa minsup=0.005 i minconf=0.4 kako bi se demonstriralo ubrzanje. Bez odsecanja, vreme izvrsenja je 5.05 sekundi dok je sa odsecanjem vreme izvrsenja 1.37 sekunda. Za vece skupove stavki bi razlika bila jos znacajnija.

Krajnji rezultat je recnik asocijativnih pravila koja zadovoljavaju argumente funkcije minsup i minconf:

