I-SUNS: Zadanie č.4

Učenie bez učiteľa

Vypracoval: Jakub Šíp

AIS ID: 91419

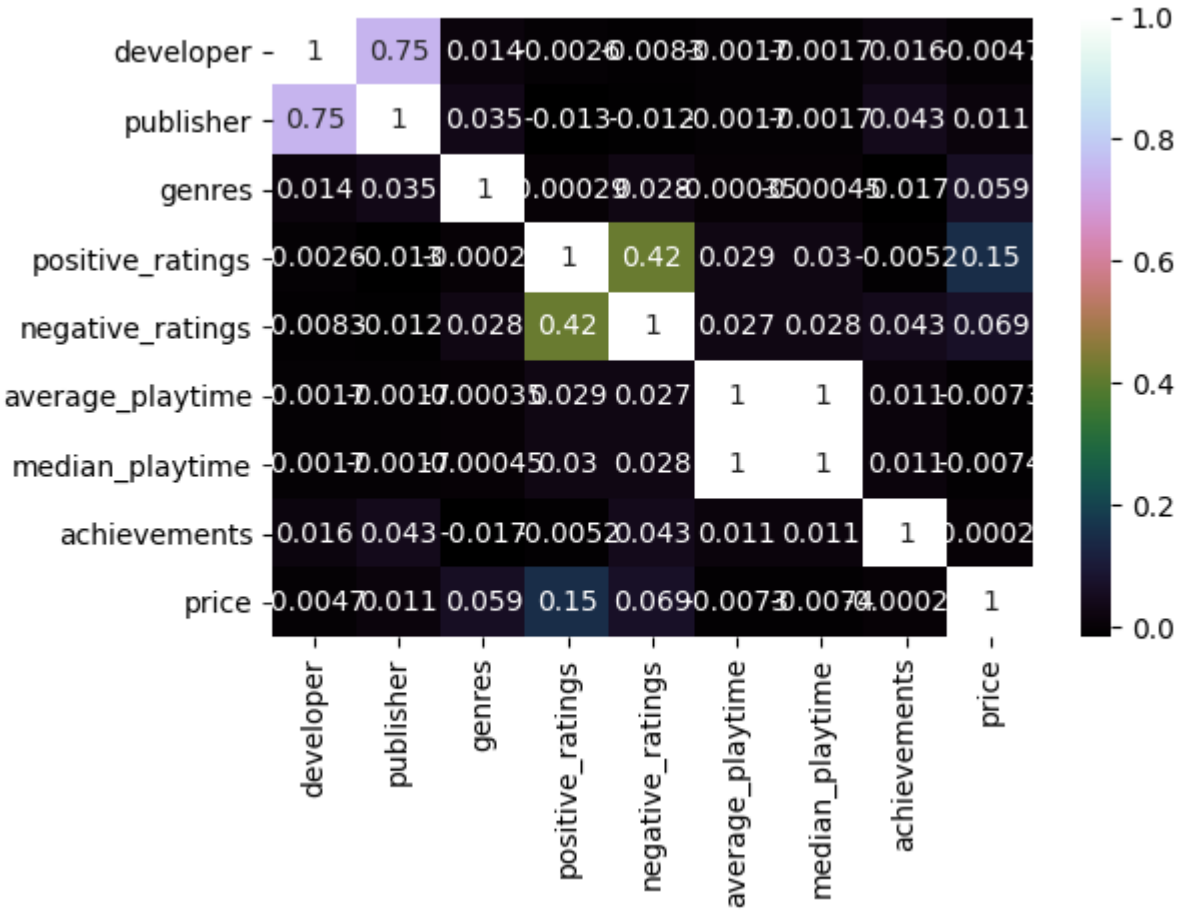
Na implementáciu tohto zadania som si zvolil programovací jazyk Python 3.8 a knižnicu sklearn. Pre predspracovanie dát som použil knižnicu pandas, ktorá uľahčuje prácu s dátovými štruktúrami. Ďalej knižnice numpy, matplotlib, seaborn

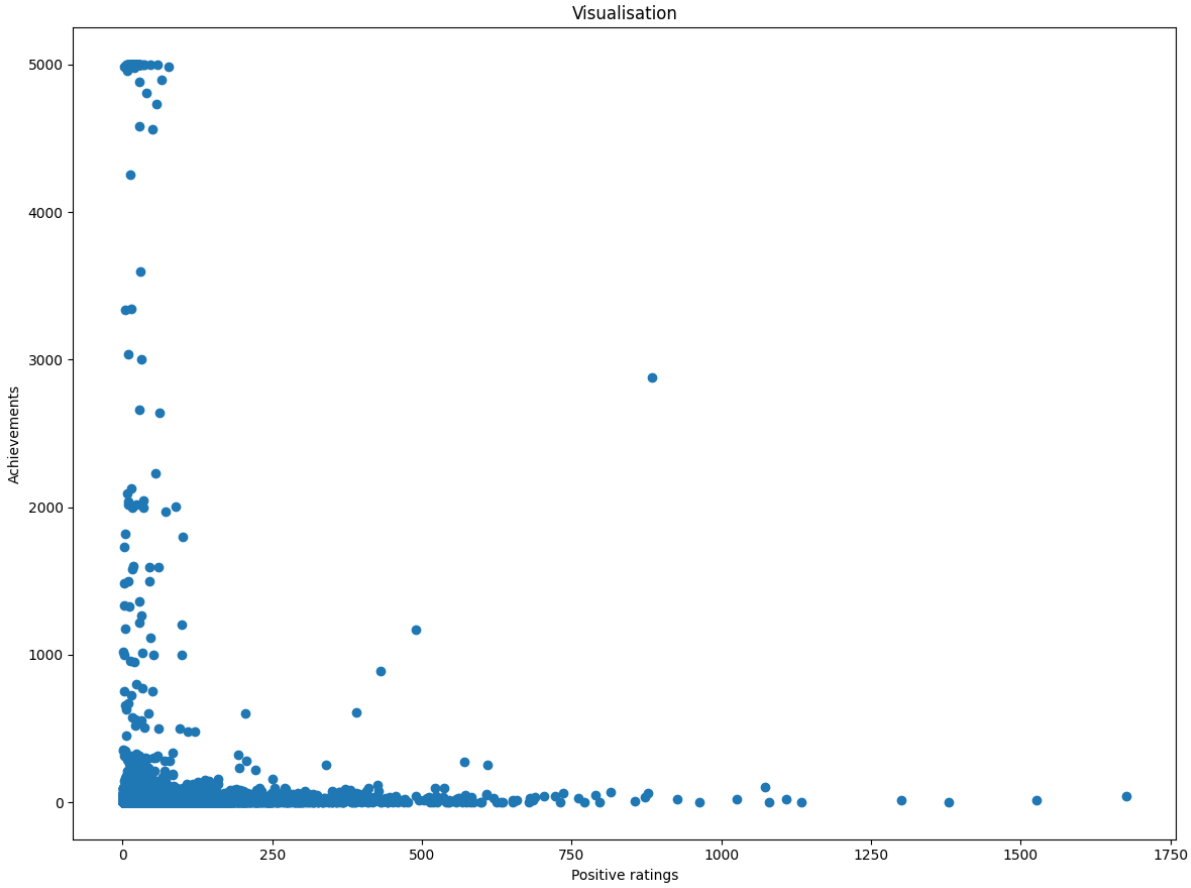
Úloha 1:

- Dáta som načítal z 2 .csv súborov. S použitím pandas spojil podľa appid . Textové hodnoty enkódoval s použitím LabelEncoder funkcie z knižnice sklearn. Následne vytriedil aby mnou použité dáta pokrývali najväčšiu časť datasetu. Takto som sa vyhol extrémom a možným chybám v datasete.

- Zmazal som dáta podľa týchto kritérií
  - negative_ratings < 35
  - positive_ratings < 20000
  - positive_ratings > 0
  - negative_ratings > 0

- Heat mapa vytvorená z môjho datasetu





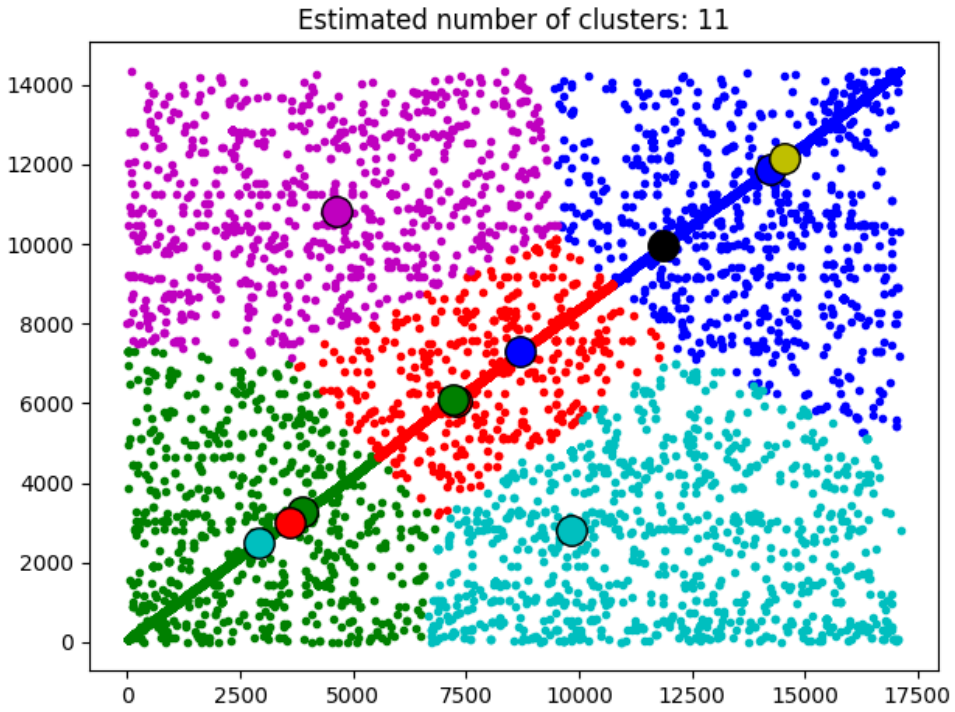
- Tu vidno, že hry rating hry nemá žiadne vplyv na množstvo achievmentov, ktoré autori do hry zapracovali a že väčšina hier má okolo 300 achievmentov a až na výnimky hráči udelia približne 500 pozitívnych hodnotení.

Úloha 2:

- Ako prvý model bez vopred určeného počtu clustrov som si zvolil algoritmus MeanShift

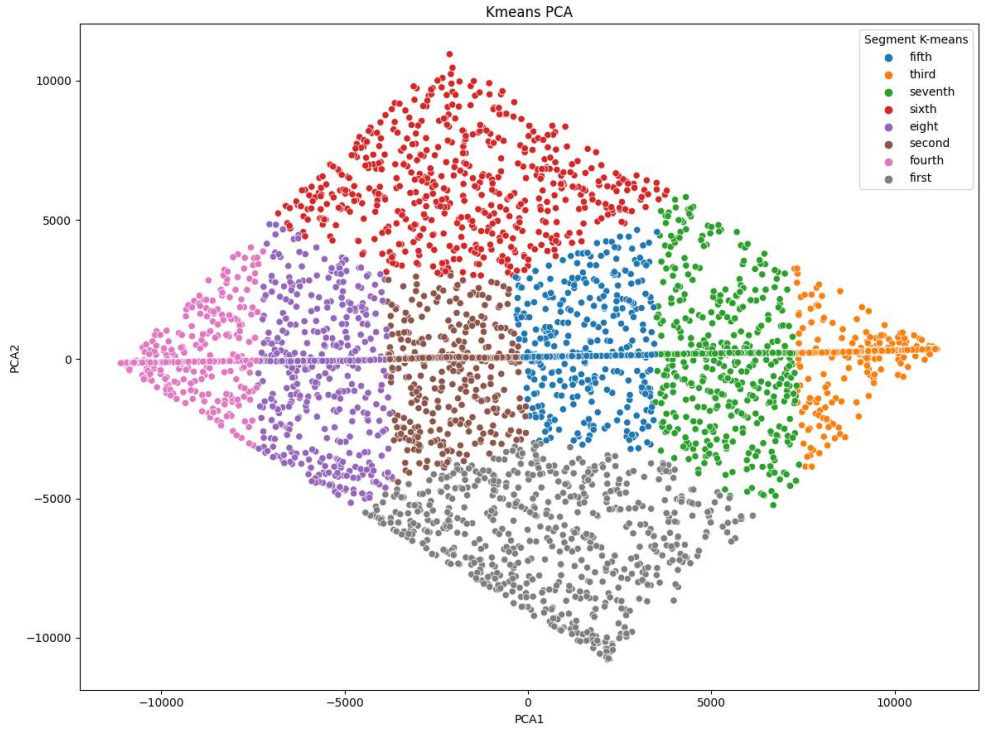
  bandwidth = estimate_bandwidth(X, quantile=0.2, n_samples=500)
  MeanShift(bandwidth=bandwidth, bin_seeding=True)

- Algoritmus odhaduje môj dataset na 11 clustrov

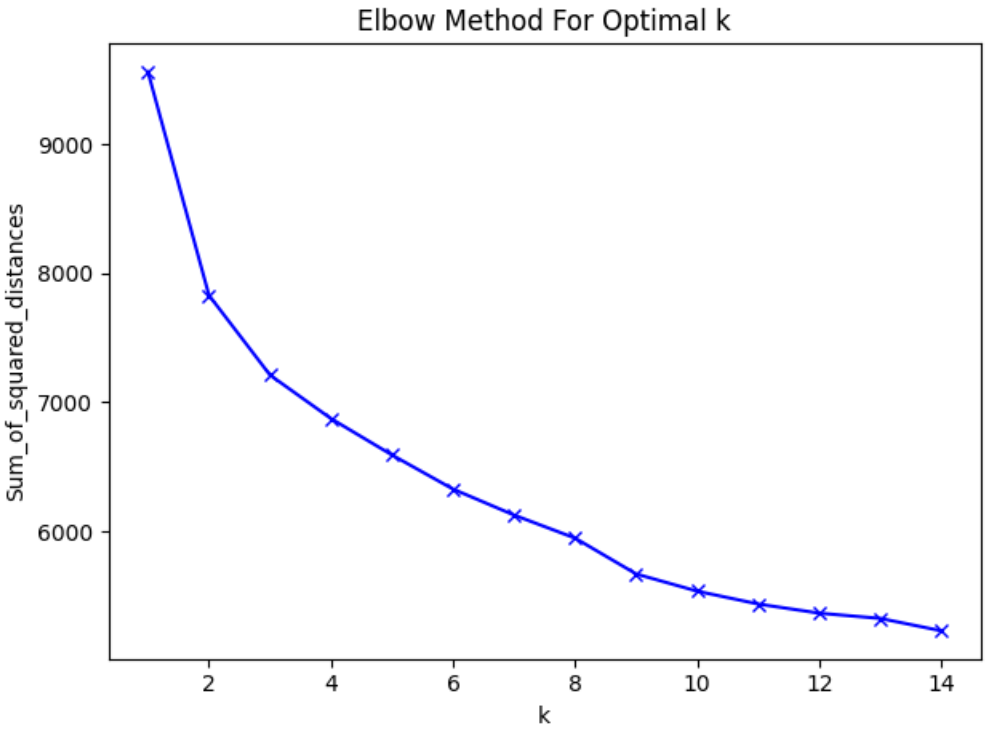


- Druhý model s vopred určeným počtom clustrov som použil Kmeans a dopredu určil počet clustrov na 8.

  KMeans(n_clusters=8, init="k-means++")



- Použil som už aj PCA, lebo bez neho boli tie dáta príliš „rozhádzane“ aj mimo svojich clustrov
- Počet clustrov som zvoli na základe tejto krivky a vybral bod kde sa krivka začala lámať (približne v strede)



Úloha 4:

- Myslíte, že takéto zhlukovanie by pomohlo pri vytvorení napr. doporučovacieho systému?

  Áno, zákazník by zadal svoje požiadavky a na základe tých by sa mu zobrazovali produkty z daného clustra. Čím bližšie by boli k bodu zákazníka tým je väčšia pravdepodobnosť, že by s takýmto návrhom bol spokojný.

- Koľko vzniká outlierov/šumu pri vašom zhlukovaní ? Aké sú to hry; prečo si myslíme, že je tomu tak?

  Po použití PCA sa tento jav už nevyskytuje, ale pred použitím PCA to bolo spôsobené pravdepodobne dátami, ktoré obsahovali napr. viac žánrov a tak zasahovali do dát z iného clustra.