Zadanie 4

Spracovanie dát

- Prvým krokom bolo spojenie dát obsahujúcich dáta o hrách s tagmi aké im užívatelia priradili. Dáta sa spájali na základe stĺpca *appid.*

- Stĺpec *developer* som rozdelil do štyroch kategórií, podľa počtu vydaných hier.
    - Kategória 1: Vydali len jednu hru.
    - Kategória 2: Vydali 2 – 5 hier.
    - Kategórie 3: Vydali 5 – 10 hier.
    - Kategória 4: Vydali 10 a viac hier.
- Stĺpec *platforms* som premapoval na tri stĺpce s binárnymi hodnotami.
- Stĺpec *genres* som premapoval na 11 rôznych kategórií. Jednotlivé kategórie je možné vidieť na obrázku:

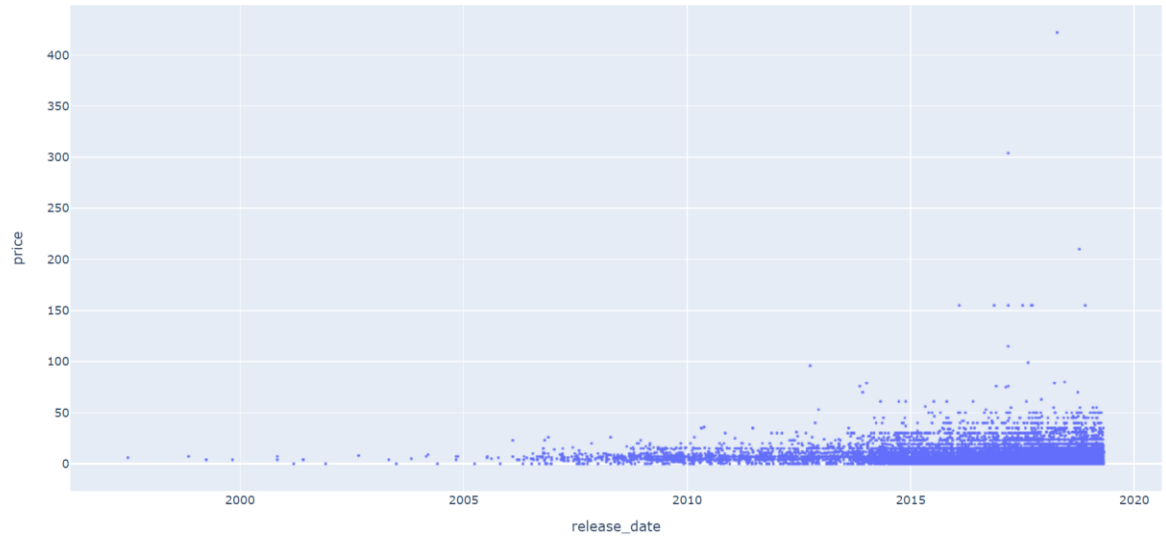
```
def get_genre(genre):
    #  if (contains(genre, "Action") == 1):
    #      return 0
      if (contains(genre, "Strategy") == 1):
          return 1
      if (contains(genre, "Casual") == 1):
          return 2
      if (contains(genre, "Indie") == 1):
          return 3
      if (contains(genre, "RPG") == 1):
          return 4
      if (contains(genre, "Adventure") == 1):
          return 5
      if (contains(genre, "Simulation") == 1):
          return 6
      if (contains(genre, "Sexual Content") == 1):
          return 7
      if (contains(genre, "Free to Play") == 1):
          return 8
      if (contains(genre, "Sports") == 1):
          return 9

      return 10
```

Kategóriu 0 som neskôr odstránil pretože sa ukázalo že je pri veľkom percente hier a to deformavalo ostatné kategórie.

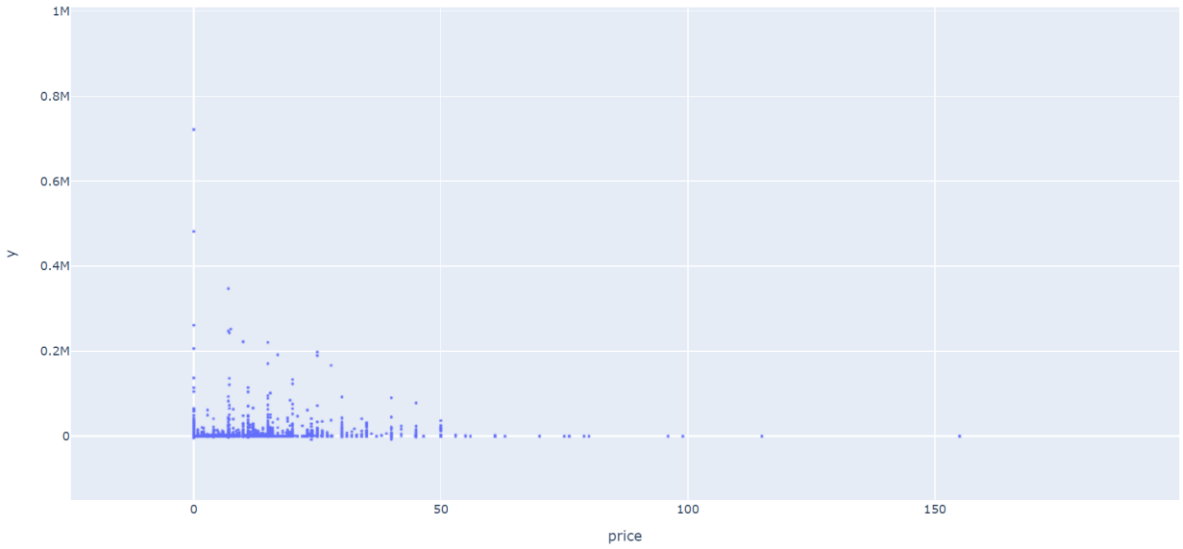
EDA

- Aký je trend vývoja cien hier?



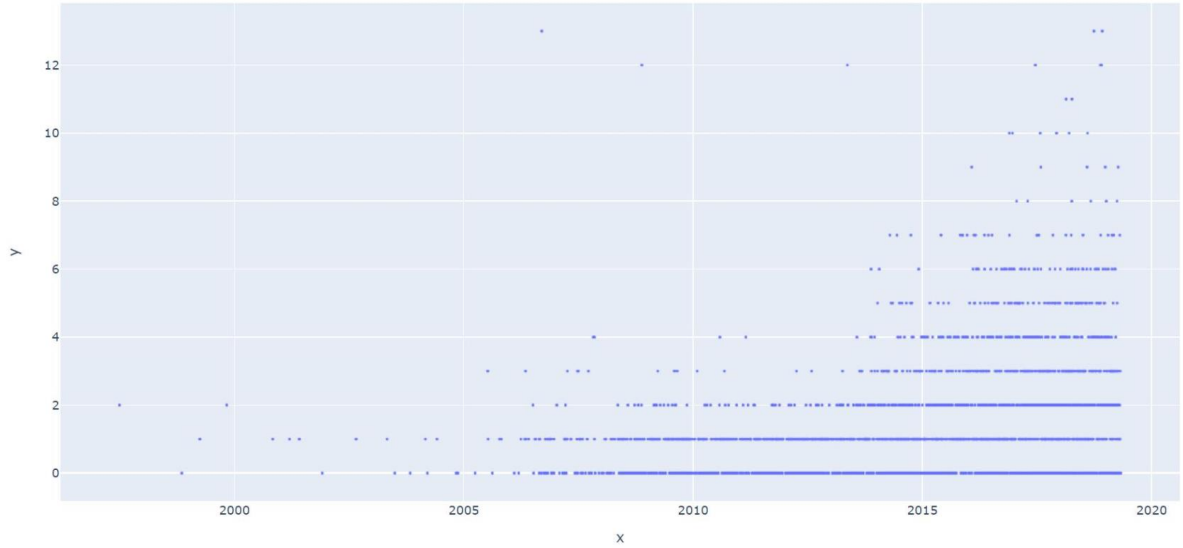
Na x-ovej osi sa nachádza rok, a na y cena hier. Z grafu môžeme jasne pozorovať že ceny hier stúpajú viac-menej lineárne. Zatiaľ čo okolo roku 2010 maximálna cena dosahovala hry niečo okolo 35 eur za hru, tak po roku 2015 sa ceny zvyšovali. A cena okolo 30 eur bola pomerne bežná. Taktiež sa významne zvýšil aj počet outlierov.

- Majú drahšie hry lepšie hodnotenie?



Na x-ovej osi je cena. Na y je nasledujúca vypočítaná hodnota: *positive_ratings - negative_ratings* pre každú hru. Na grafe nevidíme žiadny trend čo sa týka ceny a pozitívneho hodnotenia. Čím viac doprava sa na grafe nachádzame tým je pozitívnych hodnotení menej. Tu však treba zobrať aj do úvahy fakt, že drahšie majú menej hráčov a tým pádom aj menej hodnotení.

- Aký je trend vývoja hier podporujúcich multiplayer hranie?



Na grafe sú na osi x roky a na osi y počet vydaných hier v daný deň. Z grafu môžeme na prvý pohľad odčítať že pribúdajúcimi rokmi stúpal aj počet vydaných multiplayer hier. Čo je aj pravda, avšak počet hier sa všeobecne zvyšoval. Preto nieje môžné určiť či sa počet multiplayer hier vzhľadom na singleplayer hry zvyšoval.

Clustering

- Model bez vopred určeného počtu zhlukov
    - Použil som metódu DBSCAN. Hľadanie zhlukov na celých dátach trvalo pomerne dlho. Pre jedno zbehnutie to bolo približne 15 minút. Prvý krát som to pustil s default hyperparametrami. Vzdialenost 0.5 s minimálny počet vzoriek 5. Po dokončení zhlukovania však vzniklo takmer 400 rôznych klastrov. To bolo príliš veľa. Zväčšil som teda vzdialenosť na 2 a počet vzoriek na 7. Skončilo to podobným výsledkom. Vzniklo približne 200 klastrov. Testovanie som opakoval tretí krát so vzdialenosťou 5 a počtom vzoriek 15. Klustrov už vzniklo menej, okolo 30. Takýto postup som niekoľkokrát opakoval až nakoniec sa moje parametre ustálil na vzdialenosti 7 a počte vzoriek 25.

```
#    clusters = cluster.DBSCAN(eps=7,min_samples=25).fit(normalized_data)
```

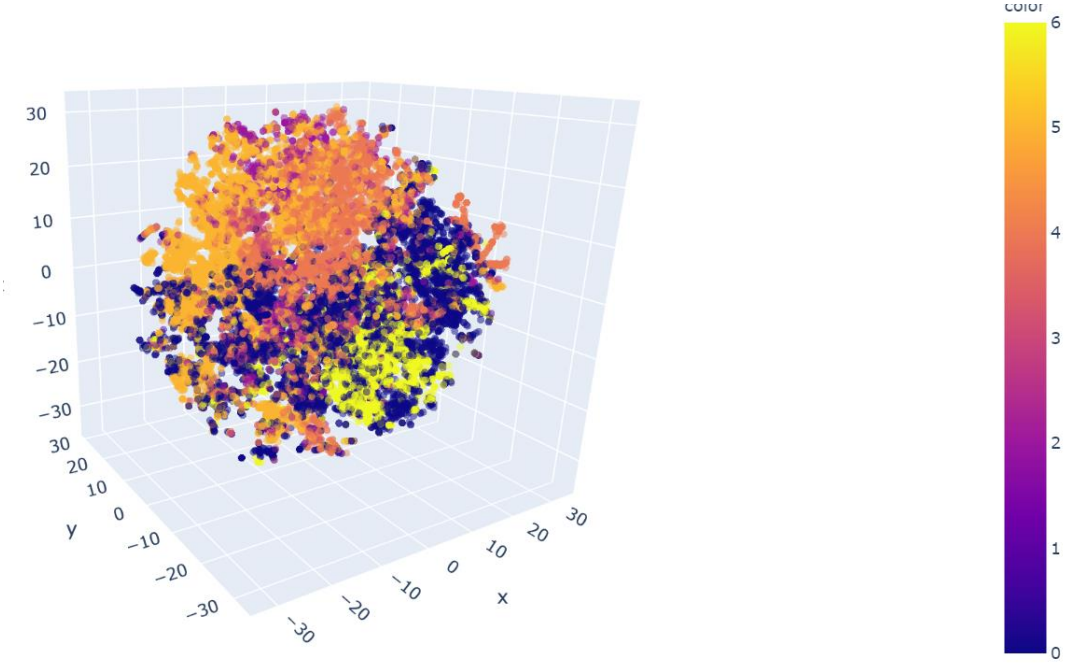
- Model s vopred určeným počtom zhlukov
    - Použil som metódu K-Means. Taktiež som vyskúšal niekoľko počtov klustrov. Nakoniec som to nechal na mojom bľubenom čísle 7.

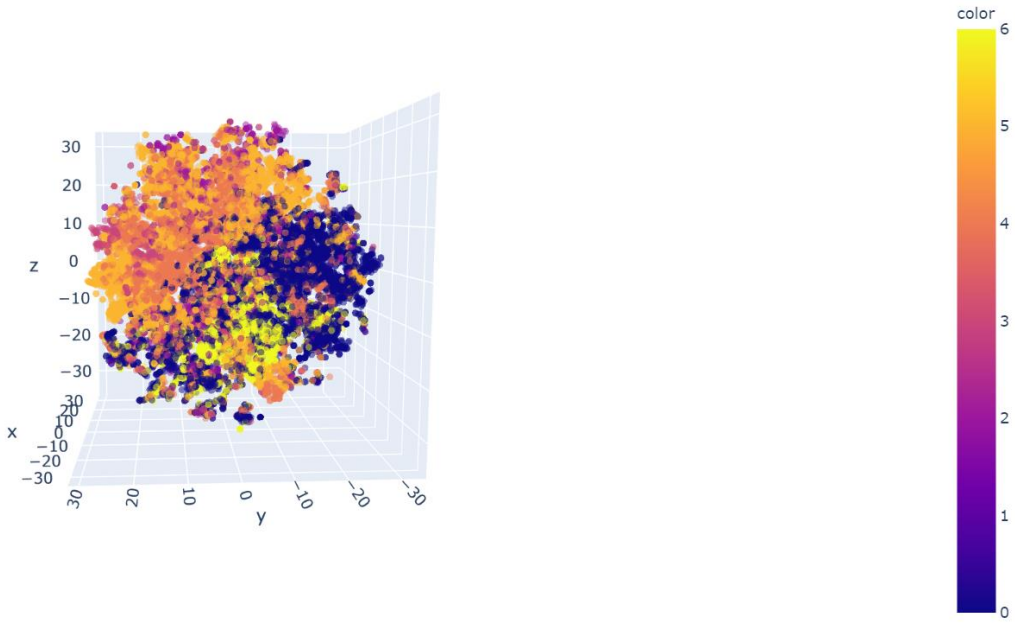
```
nClusters = 7
clusters = cluster.MiniBatchKMeans(n_clusters=nClusters, verbose=True).fit(normalized_data)
```

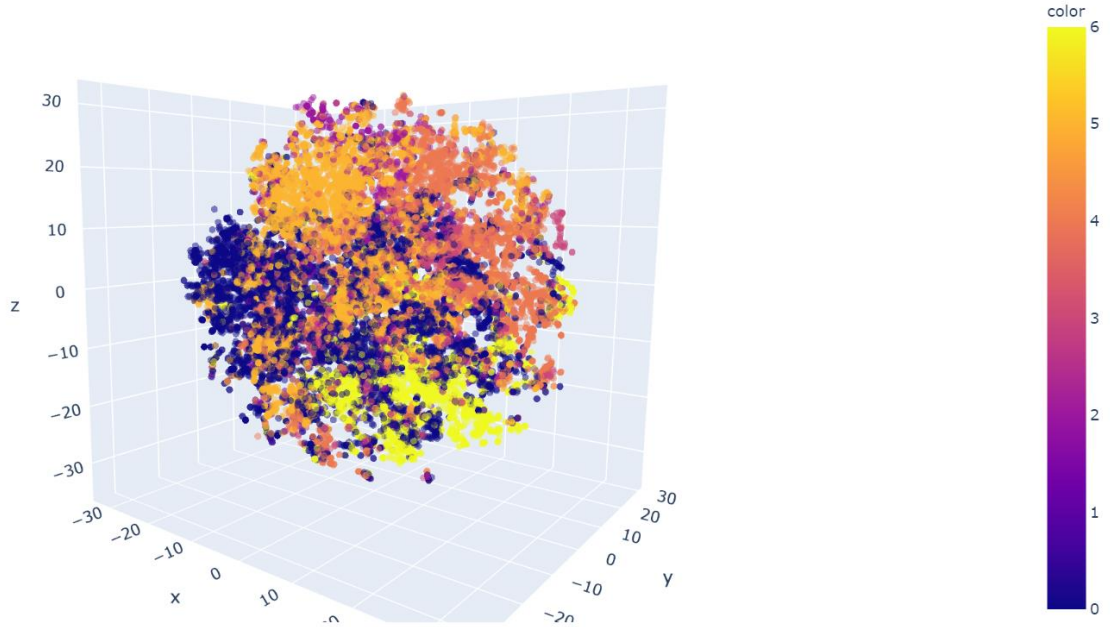
Vizualizácia clustrov

Aby som mohol dané zhluky vizualizovať, bolo potrebné zmenšiť dimenziu dát. Na to som využil dve metódy. Ako prvé som vyskúšal metódu tsne. Pomocou nej som zmenšil dimenziu dát na tri. Táto metóda však bola aj veľmi výpočtovo náročná. Môjmu PC trvalo takmer pol hodiny kým sa dopracoval k výsledku. Na obrázkoch nižšie môžeme vidieť daný výsledok vizualizovaný.

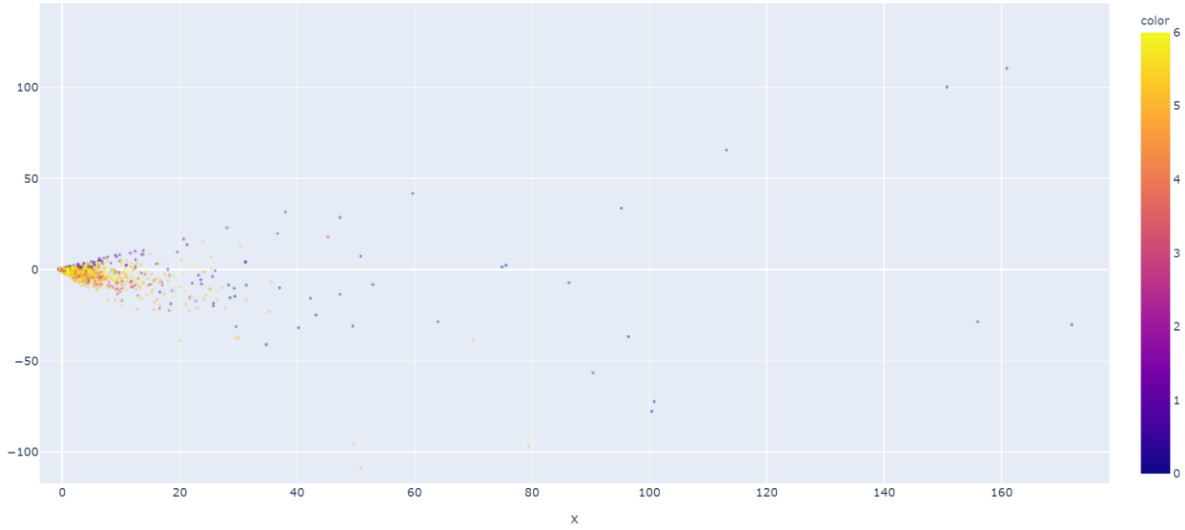
Jednotlivé clustre sú farebne rozdelené a pochádzajú z rozdelenia pomocou metódy K-Means.







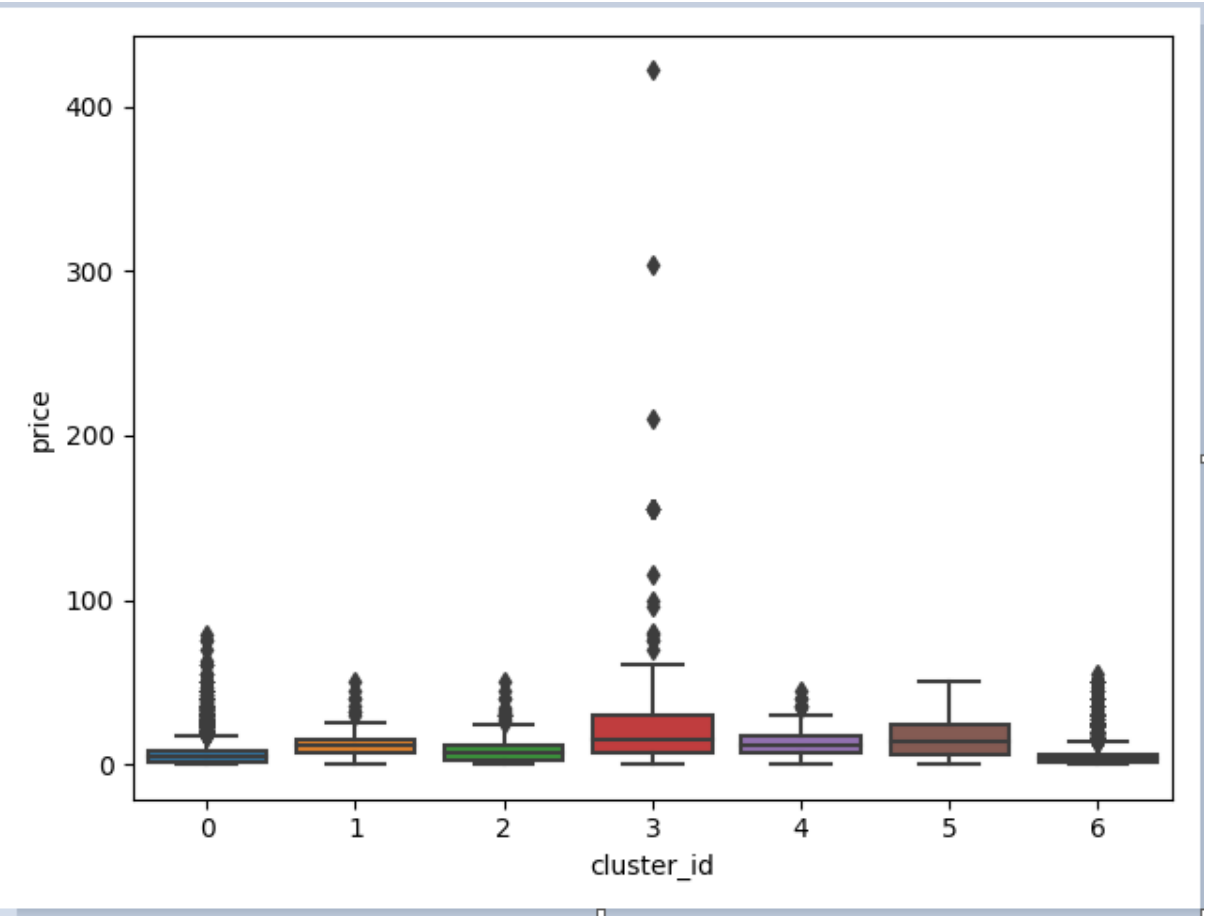
Taktiež som použil aj metódu PCA. Tu som chcel znížiť dimenziu na dva. Výsledok je možné taktiež vidieť.



Z daných dvoch obrázkov by som zhdnotil že zmenšenie dimenzie do 3d priestoru mu dalo lepšiu predstavu o dátach, ako do 2d priestoru.

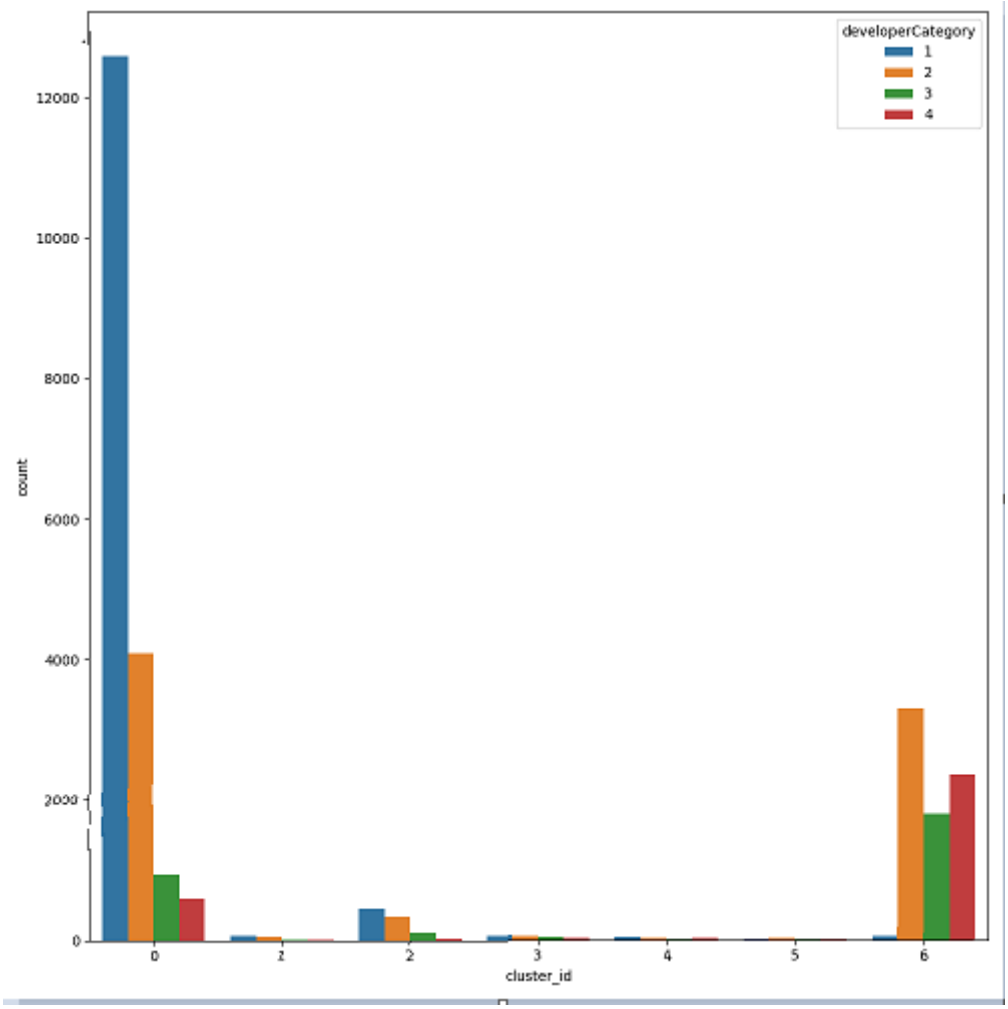
Analýza výsledkov

- Dá sa aspoň k niektorým zhlukom priradiť slovný názor/kategória?



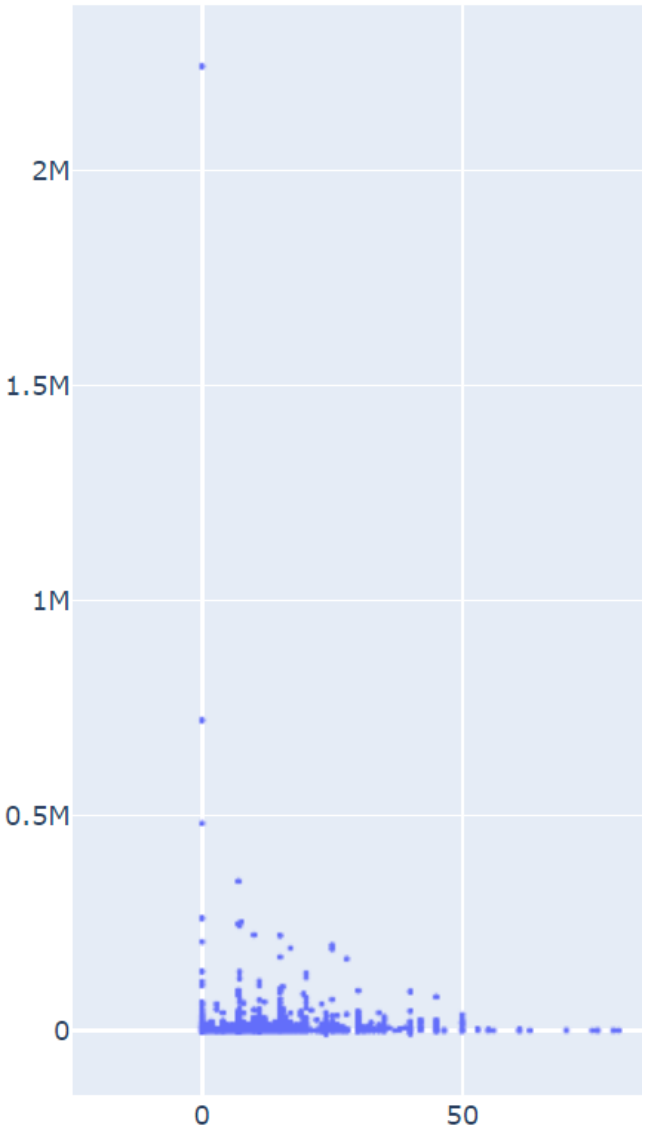
Na grafe môžeme vidieť box-plot. Na x-ovej osi sú klastre, na y sú ceny hier. Na základe tejto vizuálizácie môžeme usúdiť že v clustri 3 sú drahé hry. Nachádzajú sa v ňom tie najdrahšie hry a zároveň aj priemerná cena hier v tomto klastri je vyššia ako v ostaných.

- Ktorá hra je vhodným reprezentantom pre daný zhluk?



Na grafe vyššie môžeme vidieť zastúpenie developerských kategórií (Popísaných na začiatku dokumentácie) v jednotlivých klastroch. Je zrejmé že v klustri 0 sa nachádzajú hry od takých developerov ktorí vydali jednu až 5 hier. V klustri 6 taký ktorí vydali 5 a viac hier. Pre zhluk 0 je preto podľa mňa ideálnym reprezentantom hra *Trailmakers*, Jej cena je 19.49 eura čo ju radí do klastra 0 aj vzhľadom na cenu, a zároveň jej developerom je FlashBulb. Štúdio, ktoré vydalo len jednu hru.

- Outliers



Na tomto grafe sa nachádza cena a pozitívne hodnotenie hry. Môžeme si tu všimnúť výrazný outlier, ktorého pozitívne hodnotenie prekračuje dva milióny. Ide o hru CS:GO ktorá je jednou z najpopulárnejších hier sveta, a to zrejme spôsobuje takúto výraznú odchýlku od ostatných hier.

- Myslíte, že takéto zhlukovanie by pomohlo pri vytvorení napr. doporučovacieho systému?
    - Podľa môjho názoru áno, pomohlo ale veľmi závisi od typu dát s akými takýto systém bude pracovať.