SUNS – Zadanie 4:
Učenie bez učiteľa

Petra Kirschová

1A NAČÍTANIE A PRÍPRAVA DÁT

Na načítanie a prípravu dát som použila knižnice `pandas` a `numpy`.

Ako prvé som upravila v datasete *steam.csv* stĺpce *owners, release_date, positive_ratings* a *negative_ratings*:

- *owners*: namiesto intervalu som použila priemernú hodnotu
- *release_date*: celý dátum som nahradila iba rokom
- *positive_ratings* a *negative_ratings*: pridala som nový stĺpec *ratings*, ktorý obsahoval celkové hodnotenie vypočítané podľa https://steamdb.info/blog/steamdb-rating/:

```
sum_ratings = data["positive_ratings"] + data["negative_ratings"]
avg_ratings = data["positive_ratings"] / sum_ratings
data["ratings"] = avg_ratings - (avg_ratings - 0.5) * (2 ** -np.log10(sum_ratings + 1))
```

Z developerov a publisherov som v datasete nechala iba publisherov, nakoľko tento stĺpec obsahuje menej unikátnych hodnôt ako v prípade developerov a takisto jednotlivé hry nemajú toľko publisherov (max 5) ako developerov (max 15). Publisheri a developeri boli vo veľa prípadoch zhodní, alebo bol publisherom jeden z developerov, preto mi prišlo zahrnutie obidvoch stĺpcov zbytočné.

Pre publisherov som vytvorila 3 nové stĺpce:

- *publisher_games_count* – počet hier vydaných 1 publisherom
- *publisher_avg_owners* – priemerný počet vlastníkov hry publishera
- *publisher_avg_rating* – priemerné hodnotenie hry publishera

Žánre som spracovala tak, že som najskôr získala všetky unikátne žánre, ktoré sa v datasete nachádzali, s tým, že 1 hra mohla byť zaradená do viacerých žánrov. V základnom datasete bolo 29 žánrov. Z nich som zmazala žánre, ktoré ma v rámci analýzy až tak nezaujímali, spojila som dokopy viacero žánrov, ktoré spolu súviseli a zmenila som niektoré názvy. Po úprave zostalo 12 žánrov. Žánre som reprezentovala tak, že som vytvorila nový dataframe *genres_df* s appid a stĺpcom pre každý žáner. V datasete boli binárne hodnoty: 0, ak hra s appid nie je zaradená do žánru a 1 ak je zaradená.

K žánrom som pridala tagy zo *steamspy_tag_data.csv*. Z tagov som vybrala všetky tie, ktoré sa zhodovali zo žánrami zo základného datasetu + k nim som pridala aj ďalšie tagy.

```
genres_tags = [
    # tagy - zanre
    'action', 'rpg', 'software', 'adventure', 'indie', 'education', 'strategy',
    'simulation', 'sports', 'early_access', 'racing', 'multiplayer',
    # tagy - nove
    'singleplayer', 'arcade', 'classic', 'fantasy', 'historical', 'horror', 'puzzle',
    'sci_fi', 'shooter', 'survival', 'vr', '2d', '3d', 'addictive', 'based_on_a_novel',
    'difficult', 'great_soundtrack', 'masterpiece'
]
```

Spojila som dataframe *genres_df* s vybranými stĺpcami zo *steamspy_tag_data.csv.* Ak sa názvy stĺpcov opakovali, tak tie, ktoré sa načítali zo *steamspy_tag_data* mali pridaný suffix „_tag“.

```
genres_tags_df = genres_df.join(tag_data[genres_tags], rsuffix="_tag")
```

Pre duplicitné stĺpce (napr. *action* a *action_tag*) som postupovala takto:

a) ak mala hra počet tagnutí napr. v *action_tag* väčší ako 0, tak sa hodnota pôvodného žánra *action* prepísala z 0/1 na počet tagnutí.

```
genres_tags_df[genre] = np.where(genres_tags_df[genre + "_tag"] > 0,
                                 genres_tags_df[genre + "_tag"],
                                 genres_tags_df[genre])
```

b) Ak hra mala priradený žáner *action* (mala hodnotu 1 v stĺpci daného žánru), ale nebola tagnutá v *action_tag*, tak sa hodnota prepísala na priemer zo stĺpca *action_tag*.

```
genres_tags_df[genre] = np.where((genres_tags_df[genre] == 1) & (genres_tags_df[genre + "_tag"] == 0),
                                 int(genres_tags_df[genre + "_tag"].mean()),
                                 genres_tags_df[genre])
```

Žánre boli po úpravách reprezentované nie binárnymi hodnotami 0/1, ale počtom tagnutí.

Nakoniec som spojila *genres_tags_df* s pôvodným dataframe (*steam.csv*), a zmazala som nepotrebné stĺpce. EDA som robila s dátami bez zmazaných outlierov, pretože ma zaujímali aj také hodnoty, ktoré sa výrazne líšili od ostatných.

1B EDA

Na vykreslenie grafov som použila knižnicu `plotly`.

2.1 KORELAČNÁ MATICA

Z korelačnej matice sa dalo vyčítať viacero závislostí, ako napríklad:

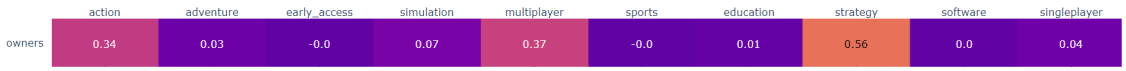
Medzi singleplayer hrami prevažujú žánre adventure, action, rpg a indie:

Multiplayer hry sú hlavne action, shooter a adventure hry:

Zaujímavá je tiež korelácia medzi žánrami a soundtrackom. Singleplayer hry majú podľa používateľov Steamu oveľa lepší soundtrack, ako multiplayer hry. Soundtrack je dobre hodnotený napríklad aj pri rpg hrách:

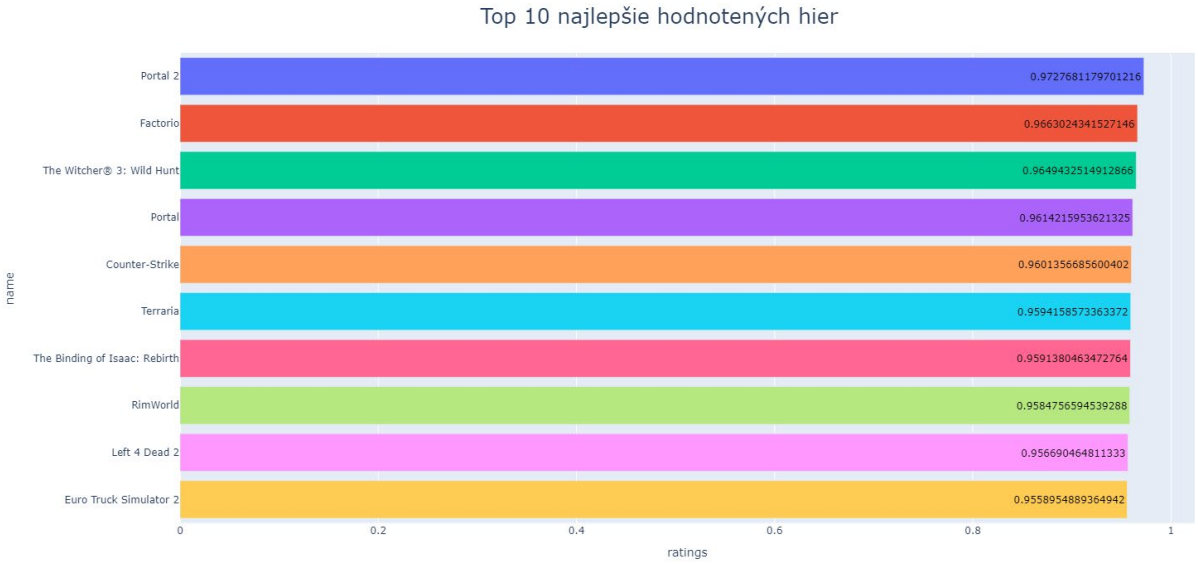
Z hier, ktoré sú na Steame dostupné ako early-access prevládajú hry žánru survival:

Najviac hráčov vlastní hry žánru action a strategy. Tiež viac používateľov kupuje multiplayer hry, ako singleplayer.

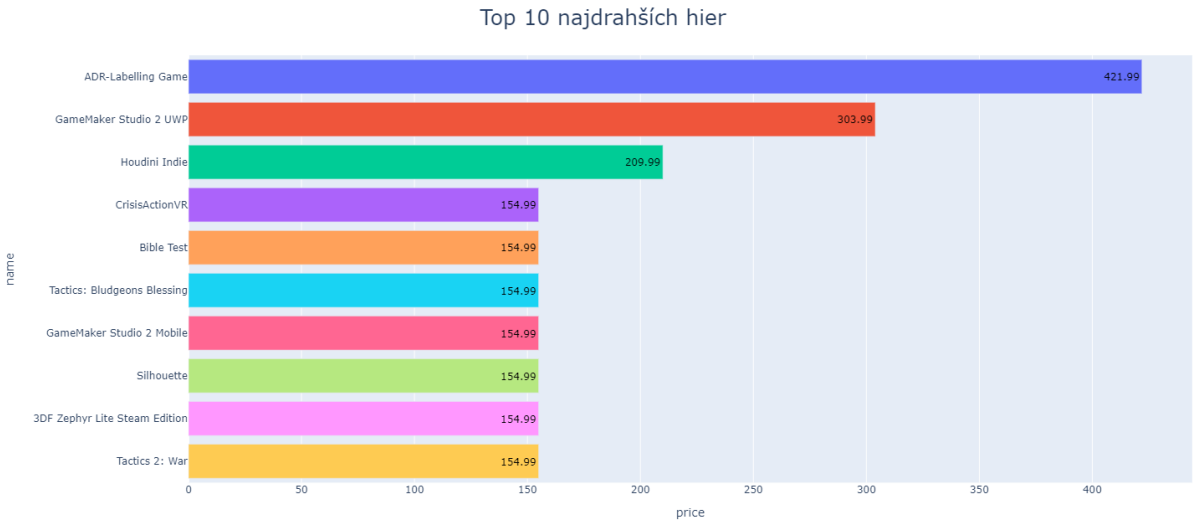


2.2 TOP 10 HIER

V top 10 najlepšie hodnotených hrách sa vyskytujú celkom známe tituly, ako napríklad Witcher 3 alebo Counter-Strike. Najlepšie hodnotenou hrou je hra Portal 2 s cca 97%.

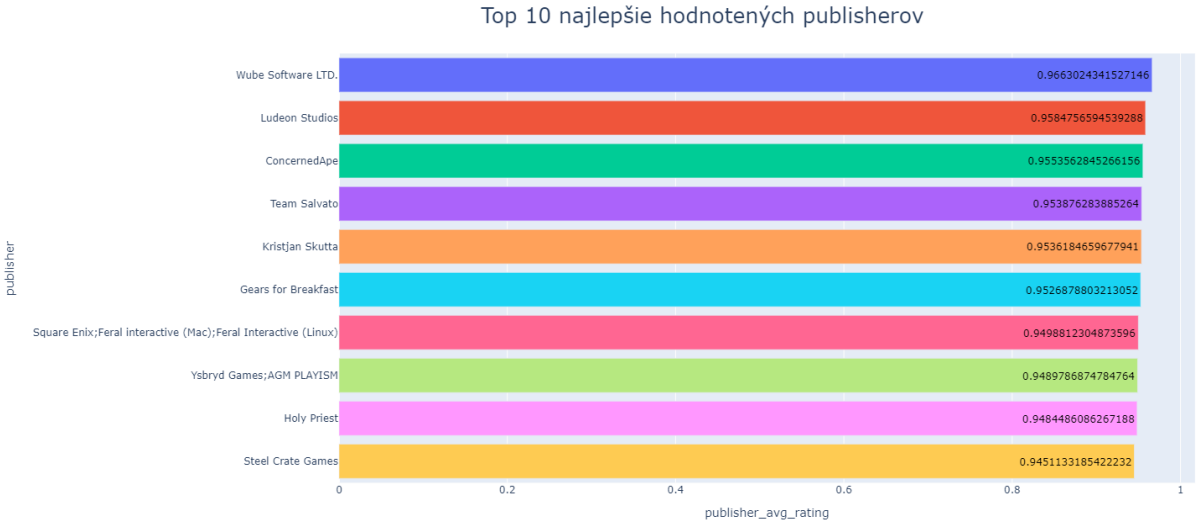


Medzi najdrahšími položkami sa vyskytujú napríklad edukačné hry a simulátory, ako ADR-Labelling Game a softvér na vývoj hier a 3D modelovanie - Houdini, GameMaker Studio.

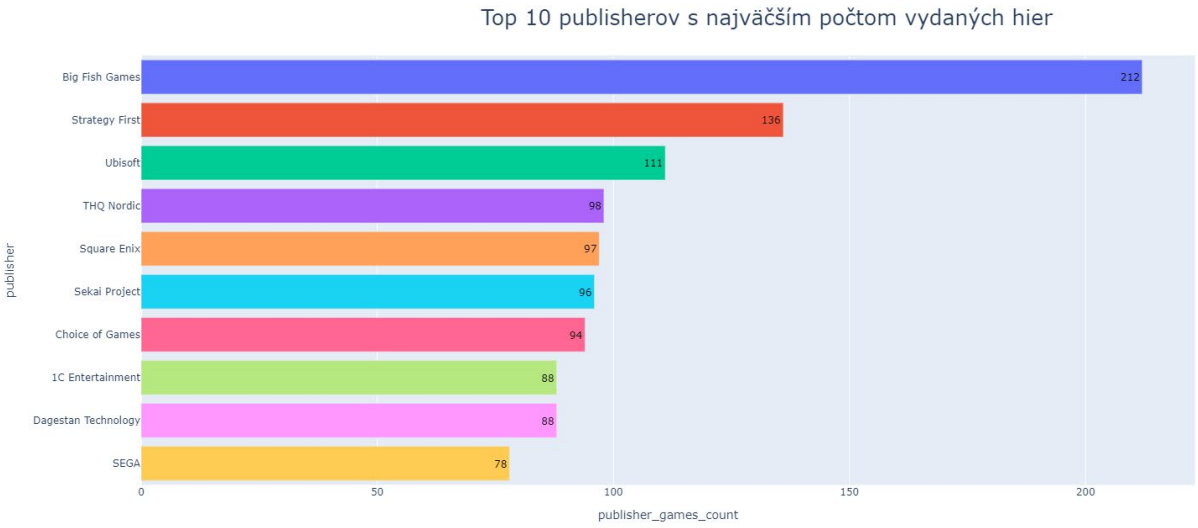


2.3 TOP 10 PUBLISHEROV

Publisheri s najlepšie hodnotenými hrami vydávajú väčšinou indie hry, ktoré sa pravdepodobne veľmi páčia ľudom, ktorí ich hrajú, no hodnotení nemajú veľa, vďaka čomu sú hodnotenia také vysoké.

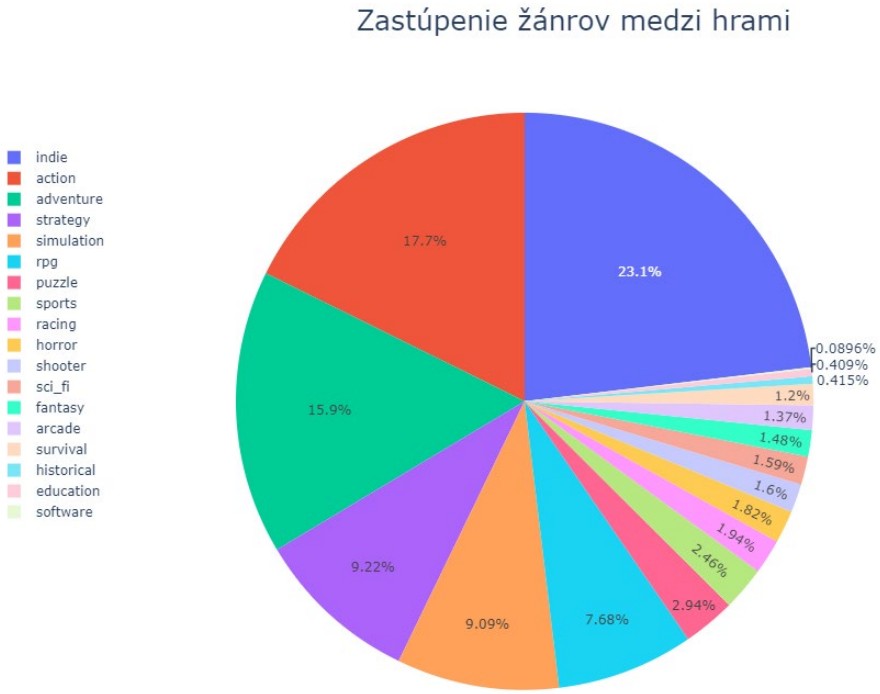


Big Fish Games prevládajú v počte vydaných hier – až 212. Sú to prevažne hlavolamy, puzzle a podobne. Najznámejší z rebríčka je Ubisoft, ktorí majú na svedomí veľa známych hier, ako Assassin's creed, Watch Dogs alebo Just Dance.



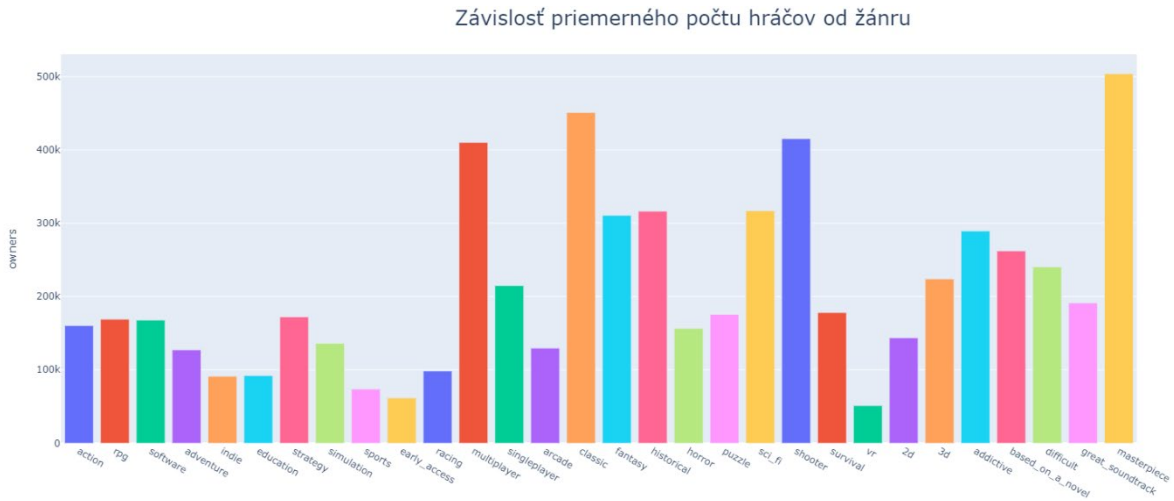
2.4 ZASTÚPENIE ŽÁNROV

Medzi žánrami prevládajú indie hry, hneď za nimi akčné a adventure hry. Najmenej patrí k rôznym druhom softvéru a tiež historickým a edukačným hrám.



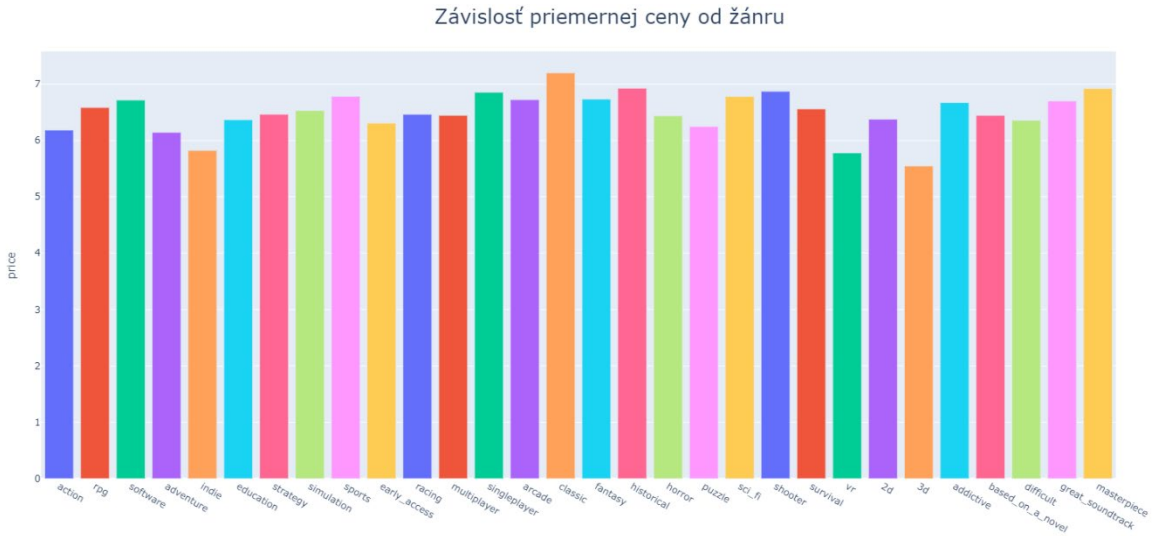
2.5 ZÁVISLOSŤ POČTU HRÁČOV OD ŽÁNRU

Hry, ktoré sú najviac predávané boli zároveň označené ľuďmi ako masterpiece a classic. Tiež veľa používateľov vlastní strieľačky a v predaji prevládajú multiplayer hry nad singleplayer. Najmenej ownerov spadá pod VR, čo môže byť čiastočne z toho dôvodu, že VR hier nie je na trhu tak veľa, ako tých z ostatných žánrov a tiež VR hry sa nedajú spustiť na akomkoľvek počítači a potrebujú VR headset, ktorý nemá každý.



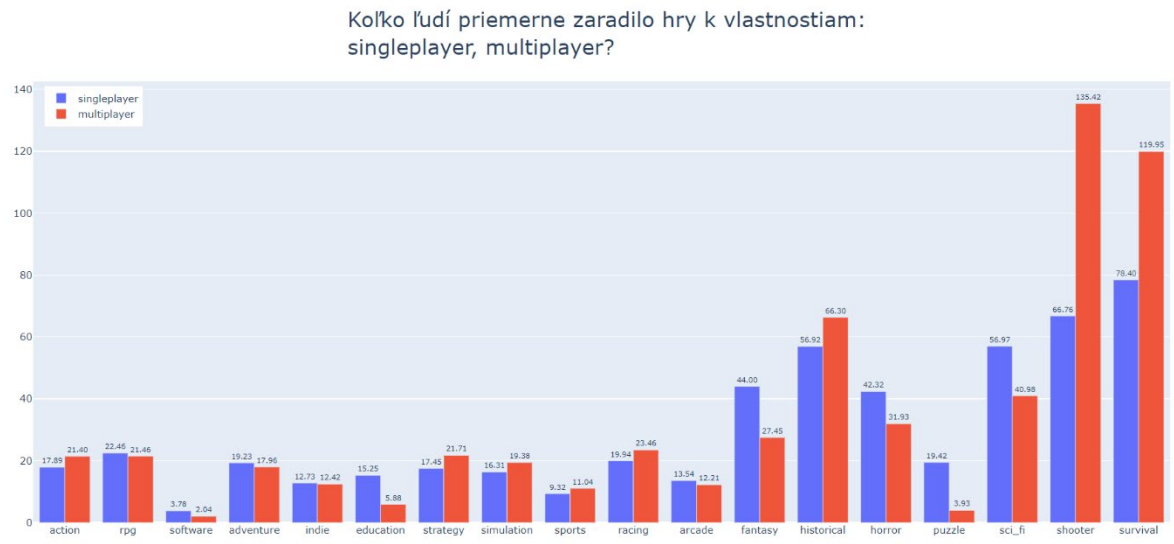
2.6 ZÁVISLOSŤ PRIEMERNEJ CENY HRY OD ŽÁNRU

Ceny pre jednotlivé kategórie sú priemerne veľmi podobné, čo môže byť spôsobené tým, že na Steame sa nachádza veľa hier, ktoré sú zadarmo, čím sa vysoké ceny vyvážia. Priemerne najdrahšie hry sú „klasiky“ a tiež strieľačky, sci-fi hry a hry označené ako masterpiece.



2.7 AKÉ ŽÁNRE PREVLÁDAJÚ MEDZI SINGLEPLAYER A MULTIPLAYER HRAMI?

Multiplayer hry vo všeobecnosti prevládajú nad singleplayer, pričom hry, ktoré boli najviac označované ako multiplayer patria medzi strieľačky a survival hry. Singleplayer hry prevládajú nad multiplayer hrami hlavne v žánroch fantasy, puzzle, horor a sci-fi.



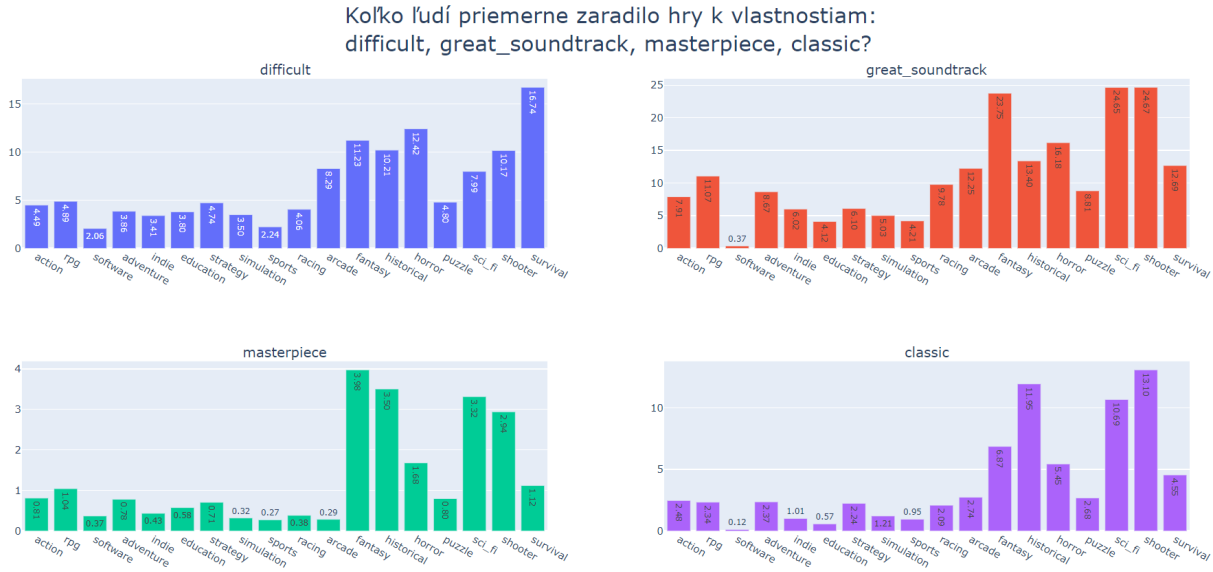
2.8 ZÁVISLOSŤ ŽÁNRU OD OSTATNÝCH VLASTNOSTÍ

Hry, ktoré sú najviac označované ako ťažké sú hlavne survival a hororové hry. Najmenej označených týmto tagom je športových hier a softvéru. Môže to byť aj z toho dôvodu, že hier v tejto kategórií bolo v datasete oveľa menej, ako ostatných.

Dobrý soundtrack majú predovšetkým fantasy, sci-fi a strieľačky. Naopak, softvéru s dobrým soundtrackom je veľmi málo (čo je pochopiteľné :)).

Ako masterpiece boli označované hlavne fantasy a historické hry, tiež sci-fi a strieľačky. Ostatným žánrom sa v tejto kategórií veľmi nedarilo.

Do klasických hier boli zaradené predovšetkým strieľačky, sci-fi a historické hry. Športových, edukačných hier a softvéru je v tejto kategórii veľmi málo.



2.9 ZÁVISÍ POČET HRÁČOV OD CENY HRY?

Zvyšovaním počtu vlastníkov klesá cena, z čoho vyplýva, že používatelia viac kupujú a hrajú hry s nižšími cenami.

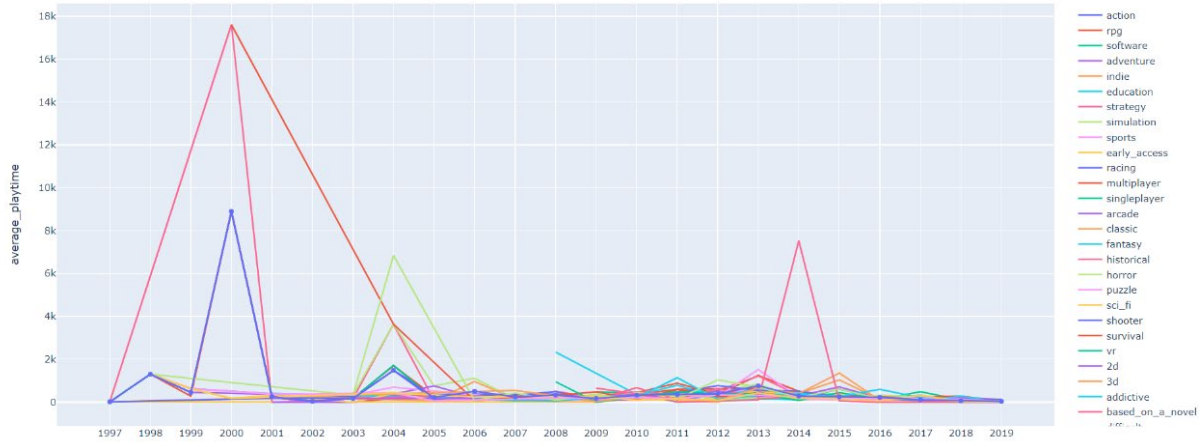
2.10 ZÁVISÍ POČET VYDANÝCH HIER OD RATINGU PUBLISHERA?

Keď publisher vydáva viac hier, jeho popularita nenarastá, naopak, hry nedosahujú také vysoké hodnotenia, ako pri publisheroch s menším počtom vydaných hier.

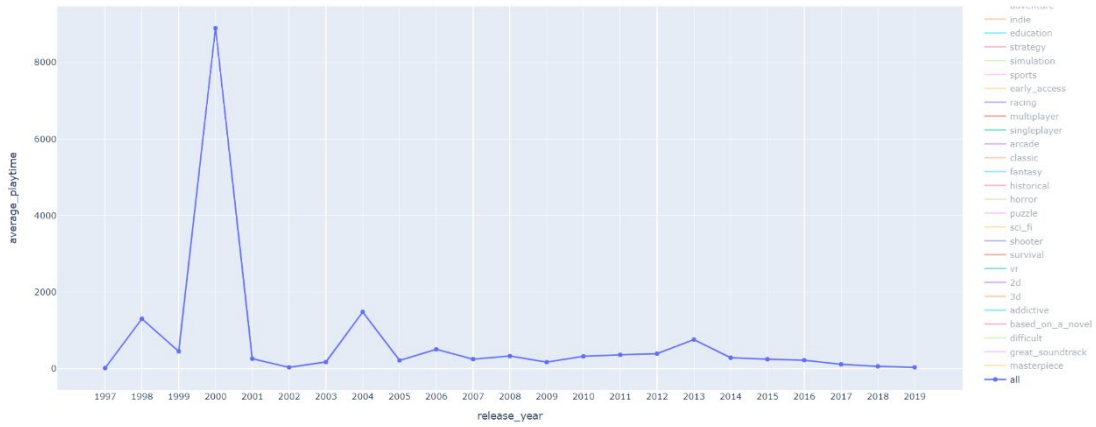
2.11 ZÁVISLOSŤ POČTU ODOHRATÝCH HODÍN OD ROKU VYDANIA

Na grafoch sú zobrazené priemerné časy hrania pre jednotlivé roky vydania pre každý žáner/kategóriu zvlášť (zobrazené na 1 grafe) aj pre všetky kategórie dokopy. Z grafov je vidno, že priemerne najviac hrané hry boli vydané okolo roku 2000. Priemer výrazne ovplyvnila veľmi populárna hra Counter-Strike, ktorá bola vydaná v roku 2000 a mala veľmi vysoký playtime.

Priemerný počet odohratých hodín – každá kategória zvlášť

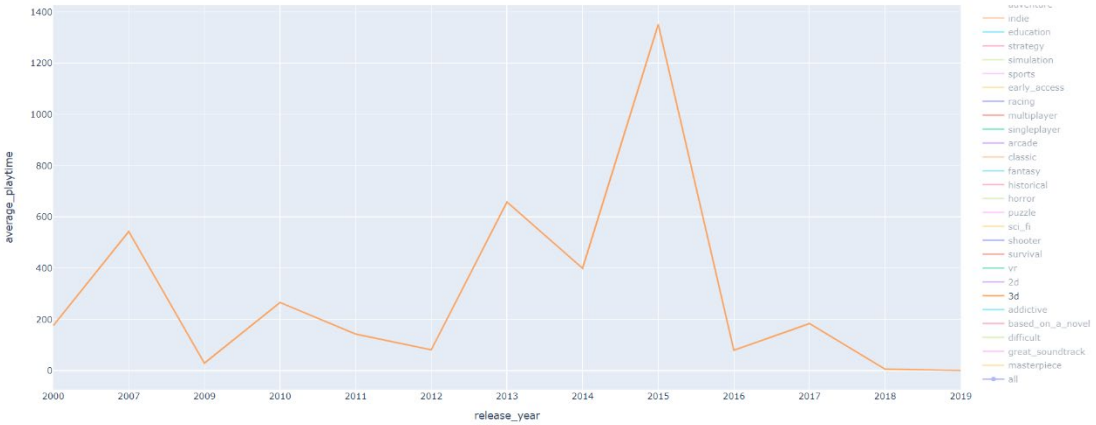


Priemerný počet odohratých hodín – všetky kategórie



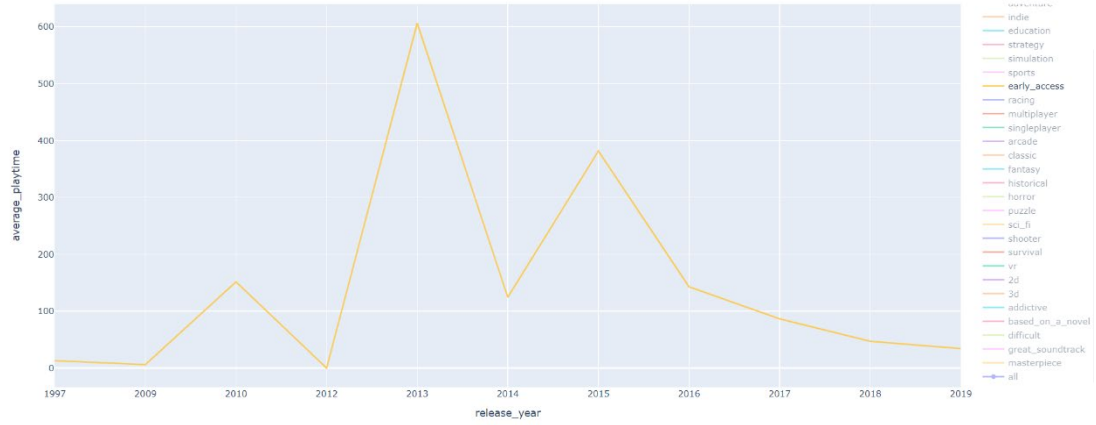
3D hry napríklad dosiahli maximum v roku 2015. V tomto roku vyšlo veľa známych a ikonických hier, ako Witcher3, GTA 5, Call of Duty: Black Ops 3, Kerbal Space program a podobne, čo sa odrazilo aj na priemerne odohratých hodinách pre hry v tomto roku:

Priemerný počet odohratých hodín - 3D



Early-access hry majú odohratých veľmi málo hodín v porovnaní s ostatnými kategóriami, čo dáva zmysel, keďže vo všeobecnosti je early-access hier na Steame menej a tiež si ich kupuje menej ľudí, ako finálne verzie hier:

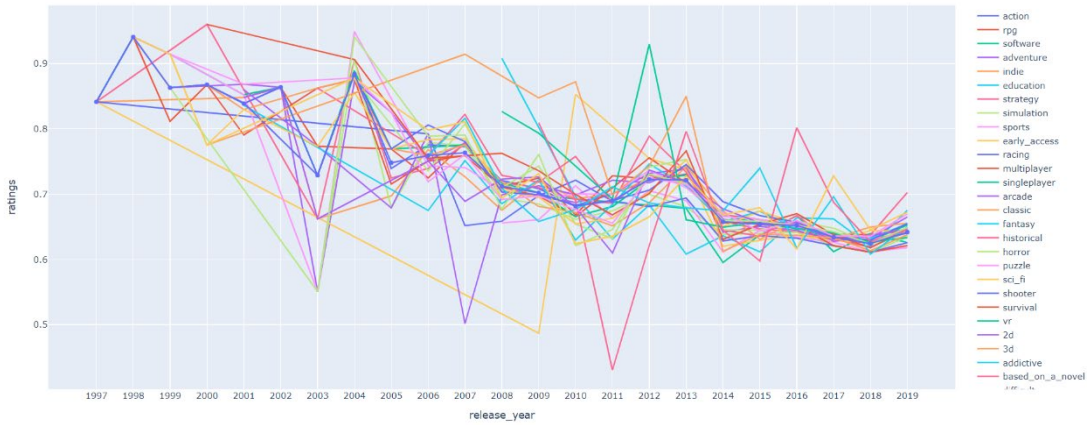
Priemerný počet odohratých hodín - Early-access



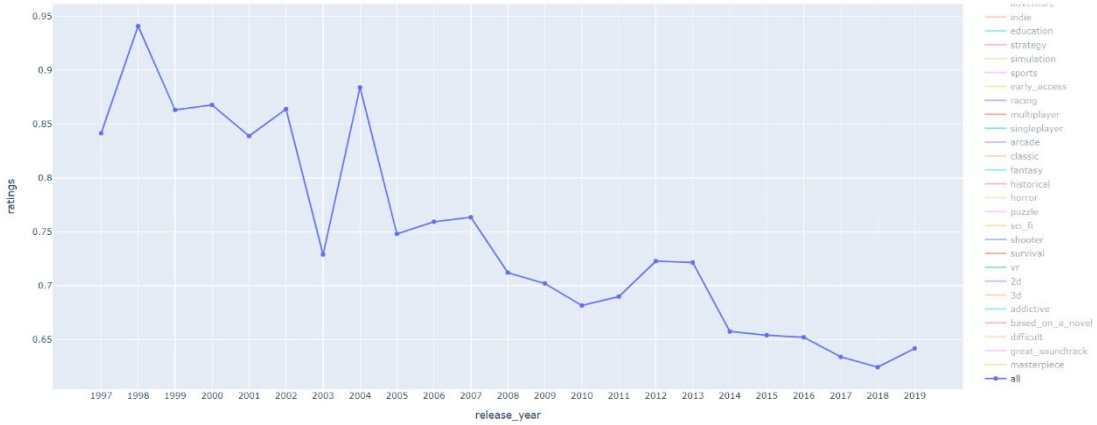
2.12 ZÁVISLOSŤ RATINGU OD ROKU VYDANIA

Staršie hry dosahujú vyššie hodnotenie, ako novšie hry, čo je vidno jednak na grafe pre jednotlivé kategórie zvlášť, ako aj pri ratingoch pre všetky žánre dokopy.

Rating – každá kategória zvlášť



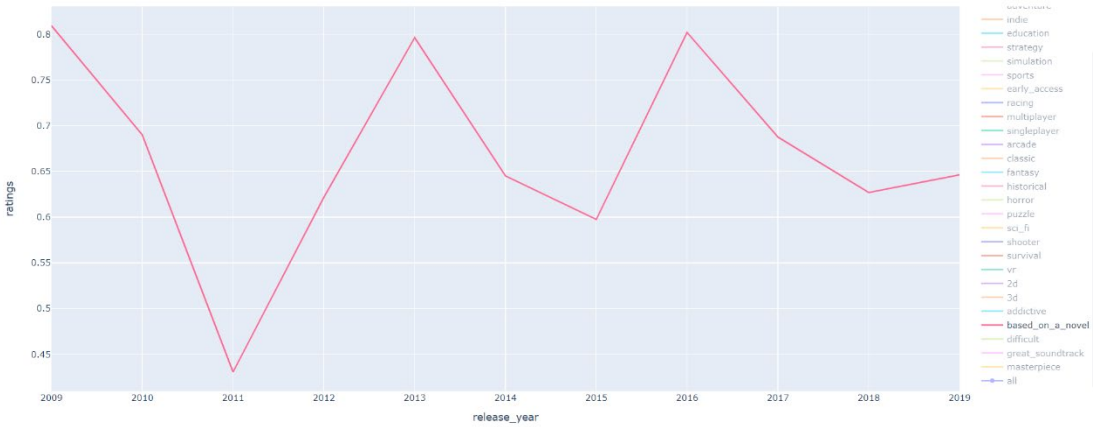
Rating – všetky kategórie



Najnižší rating získali hry založené na románoch v roku 2011. Môže to byť aj z toho dôvodu, že hier v tejto kategórií bolo celkovo dosť málo a v roku 2011 vyšla iba 1 hra s relatívne nízkym hodnotením.

name	release_year	ratings
King Arthur: Fallen Champions	2011	0.43063

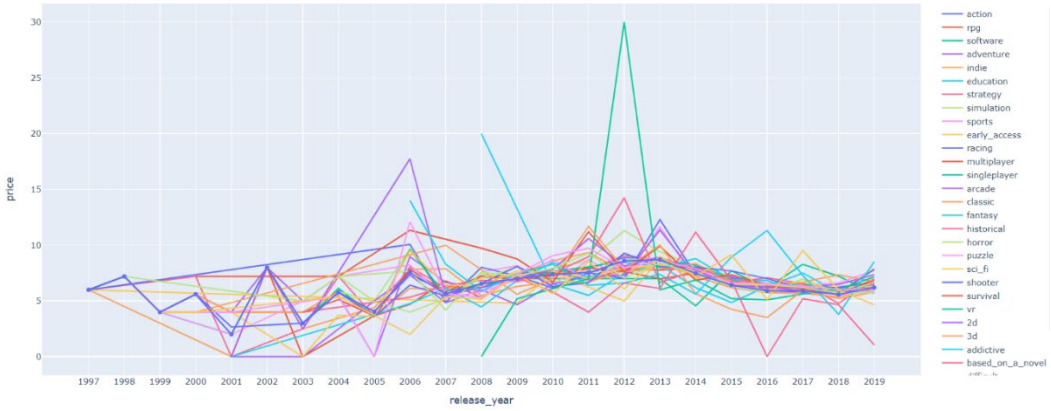
Rating – Based on a novel



2.13 ZÁVISLOSŤ CENY OD ROKU VYDANIA

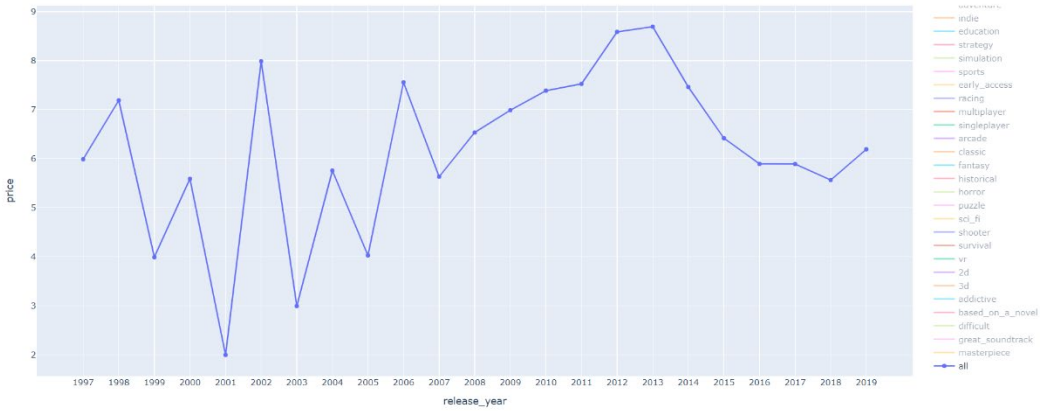
Ceny hier sú pre všetky žánre veľmi podobné, pohybujú sa priemerne v rozpätí od 0 po 15 dolárov. V niektorých rokoch je cena výrazne vyššia oproti ostatným.

Cena – každá kategória zvlášť



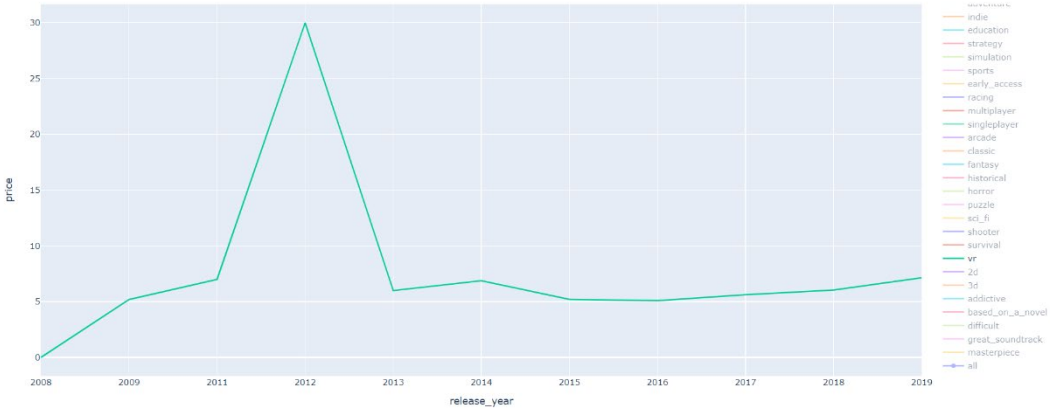
Priemerné ceny pre všetky hry spolu počas rokov dosť kolíšu, najvyššie ceny sú okolo rokov 2012 – 2013, ktoré mohli byť ovplyvnené veľmi vysokou cenou niektorej z hier.

Cena – všetky kategórie



Najväčší peak v cenách hier bol v 2012 pri VR hrách, kedy prišiel na trh prvý Oculus Rift. Ceny novších hier v priemere výrazne klesli, oproti tým z roku 2012. Časom pravdepodobne pribudli aj VR hry, ktoré boli dostupné zadarmo, čo vyvážilo prípadné vyššie ceny pri novších VR hrách.

Cena - VR



3 K-MEANS CLUSTERING

Ako model s vopred určeným počtom zhlukov som zvolila K-means. Použila som knižnicu `sklearn` a na vykreslenie grafov `plotly`.

Pred clusterovaním som vymazala outliery a dáta som štandardizovala pomocou StandardScaler().

K-means clustering som inicializovala takto:

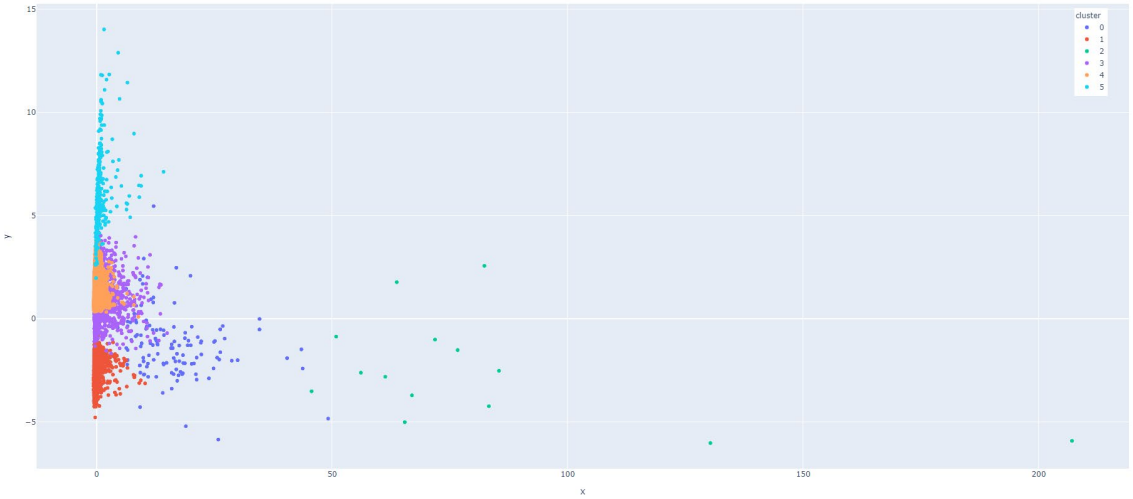
```
kmeans = KMeans(init="random",
                n_clusters=6,
                max_iter=300).fit(data_to_fit_scaled)
```

- init="**random**" – počiatočné body vyberá náhodné
- n_clusters=**6** – počet vygenerovaných clusterov
- max_iter=**300** – počet iterácií

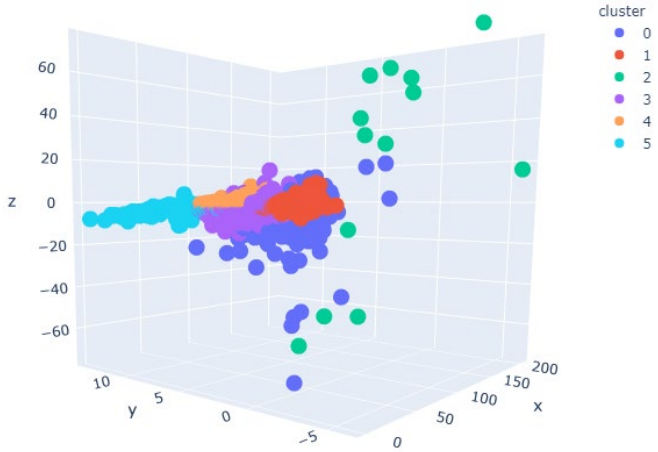
3.1 REDUKCIA DIMENZIE

Dimenziu som redukovala pomocou PCA do 2D a 3D priestoru, ktoré som vizualizovala pomocou scatter grafov. Vzorky sú na grafe blízko seba, no dá sa vyčítať, ktorý bod patrí do ktorého clustra.

PCA – 2D

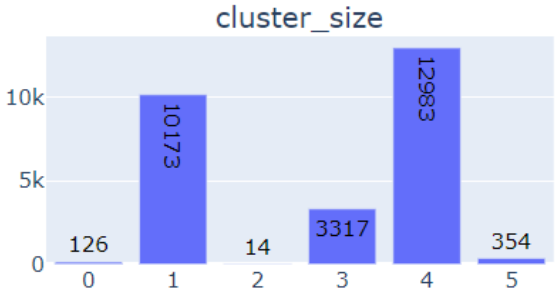


PCA – 3D



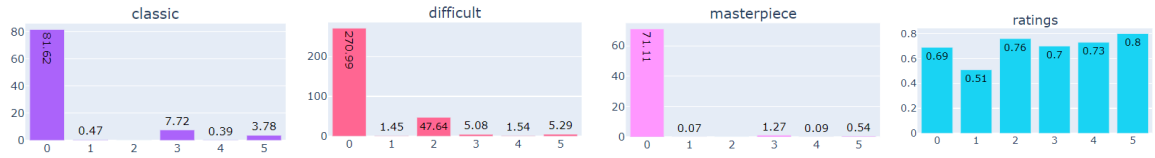
3.2 ANALÝZA

K-means vygeneroval 6 clusterov, pričom clustery s výraznejším počtom vzoriek sú 1, 3 a 4.



3.2.1 Cluster 0

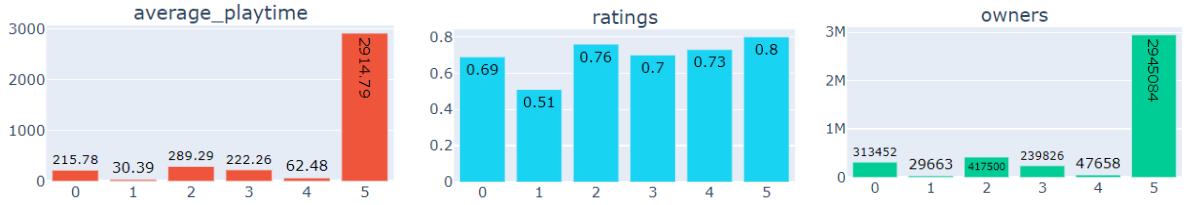
Cluster 0 je druhý najmenší cluster, prevládajú v ňom napríklad hry kategórií classic, difficult a masterpiece. Zaujímavé na tomto clustri je, že napriek tomu, že veľa ľudí označilo hry v ňom ako masterpiece a classic, priemerné hodnotenia sú paradoxne takmer najnižšie. Dalo by sa skôr očakávať, že to bude naopak – hry, ktoré sú považované za masterpiece by mali byť aj hodnotené pozitívne. Mohlo to tak byť aj z toho dôvodu, že je tento cluster malý, preto sú počty trochu skreslené.



Tento cluster by sa mohol spojiť s clusterom 1, pretože spĺňajú podobné vlastnosti – nízke hodnotenia, neznáme hry, mýlo nahraných hodín a používateľov.

3.2.2 Cluster 1

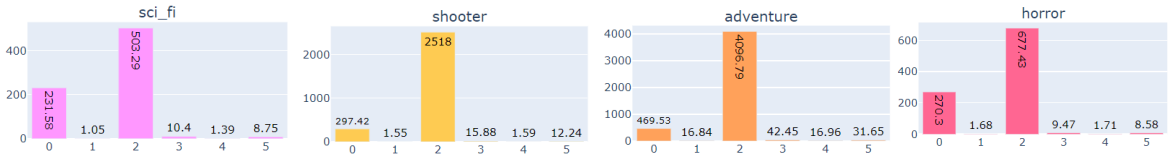
Cluster 1 obsahuje hry, ktoré majú najmenší priemerný počet nahratých hodín, najmenší počet vlastníkov a najnižšie ratingy. Dalo by sa povedať, že tento cluster reprezentuje menej obľúbené hry. Je tu mix rôznych žánrov, žiadny z nich neprevládal.



Príklady hier z clustra 1: nízky rating, neznáme hry, rôzne žánre:

3.2.3 Cluster 2

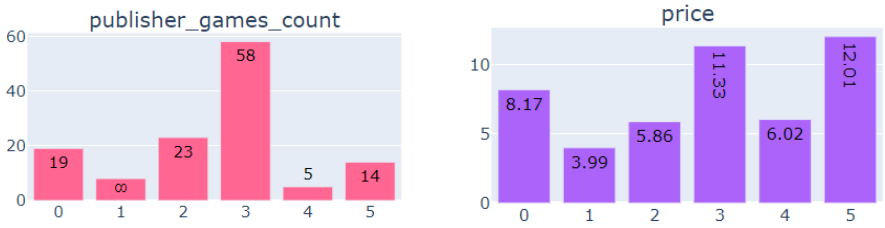
Cluster 2 bol najmenší, obsahoval iba 14 hier, no dosahoval vysoké čísla takmer pre každý žáner. Zhluk pravdepodobne obsahuje nejakého reprezentanta z každého žánru. Hry v tomto clusteri sa nedajú špecifikovať jedinou výraznejšou vlastnosťou.



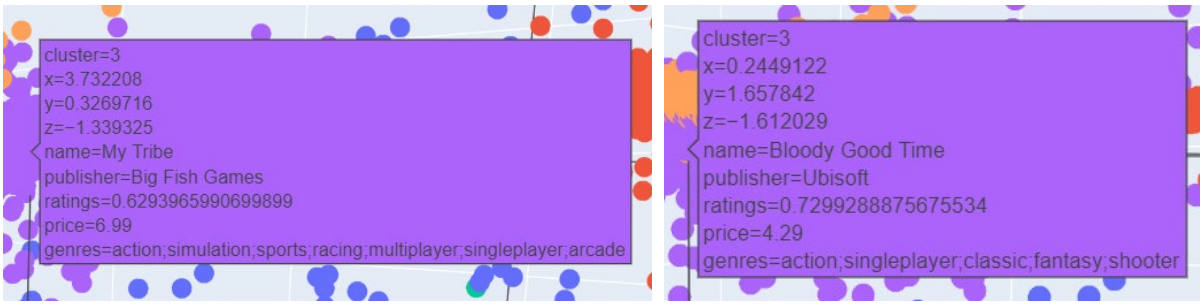
Na PCA grafoch je tento cluster celkom rozptýlený a oddelený od ostatných (je vyfarbený zelenou farbou). To môže naznačovať aj to, že tieto hry patria k šumu a nedajú sa zaradiť k žiadnemu inému clusteru.

3.2.4 Cluster 3

V clustri 3 vynikali hlavne atribúty publisher_games_count (počet hier, ktoré publisher vydal) a cena. Tento cluster by som nazvala ako hry od aktívnych vydavateľov.



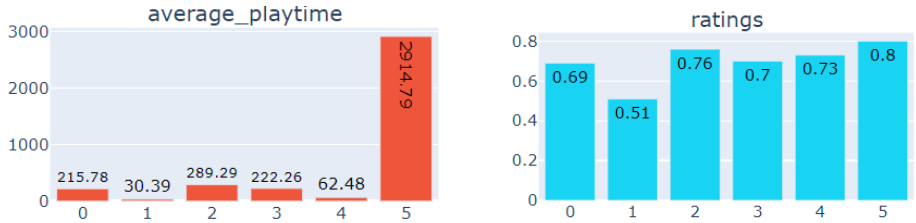
Príklady hier z clustra 3 - známi publisheri s veľkým počtom vydaných hier:



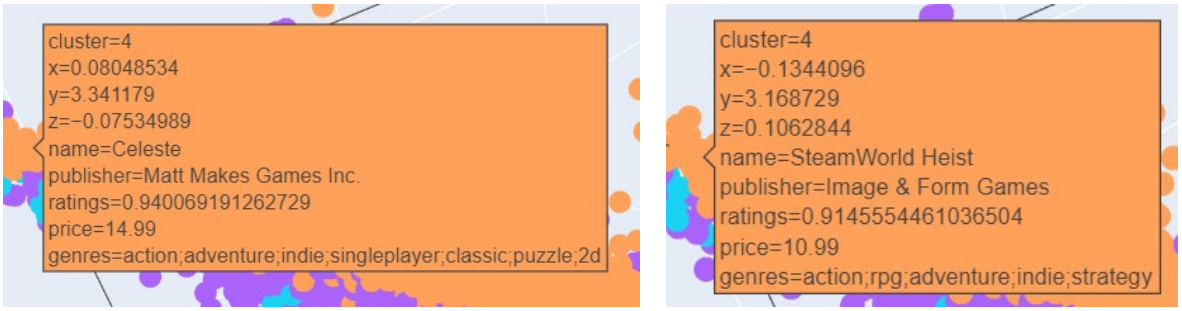
Obaja publisheri – Big Fish Games a Ubisoft patrili medzi top 10 publisherov s najväčším počtom vydaných hier.

3.2.5 Cluster 4

Cluster 4 je najväčší, no žiadny žáner v ňom neprevládal. Patria sem hry, ktoré sú málo hrané, avšak sú v priemere dosť dobre hodnotené. Označila by som ho ako menej známe, ale kvalitné hry.

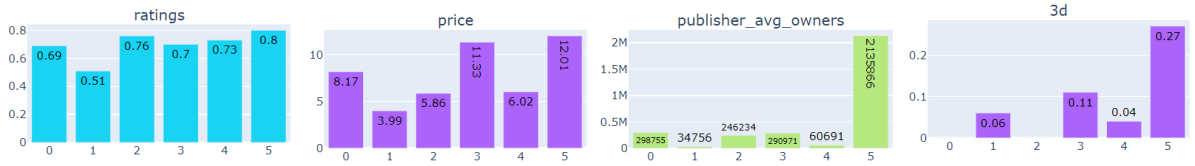


Príklady hier z clustra 4 – vysoké hodnotenia, neznáme hry:



3.2.6 Cluster 5

V clustri 5 sa nachádzajú najlepšie hodnotené a zároveň najdrahšie hry a hry s najväčším počtom vlastníkov. Z kategórií hier sú tu zastúpené hlavne 3D hry. Tento cluster by som označila ako najznámejšie a najpopulárnejšie hry.



Príklady hier z clustra 5 – veľmi známe a populárne hry:

3.2.7 Záver

„Myslíte, že takéto zhlukovanie by pomohlo pri vytvorení napr. doporučovacieho systému? (Môžete čerpať z vlastných skúseností)“

Tento model by sa mohol podľa mňa kľudne použiť v marketingu, keďže sa dá vyčítať, ktoré skupiny hier majú najvyšší počet vlastníkov a na základe toho napríklad zamerať reklamy práve na tieto kategórie. Pri tomto modeli som nemala v žiadnom clustri prevládajúcu kategóriu alebo žáner hry, preto by bolo podľa mňa ťažké spraviť nejaké doporučovanie na základe tematiky hry, skôr by sa dalo doporučovať na základe popularity, počtu predaných hier a podobne.

4 DBSCAN CLUSTERING

Pri modeli bez vopred určeného počtu zhlukov – DBSCANe som postupovala rovnako, ako pri K-means. Použila som dáta s vymazanými outliermi, ktoré som pred zavolaním funkcie štandardizovala pomocou StandardScaler().

DBSCAN clustering som inicializovala takto:

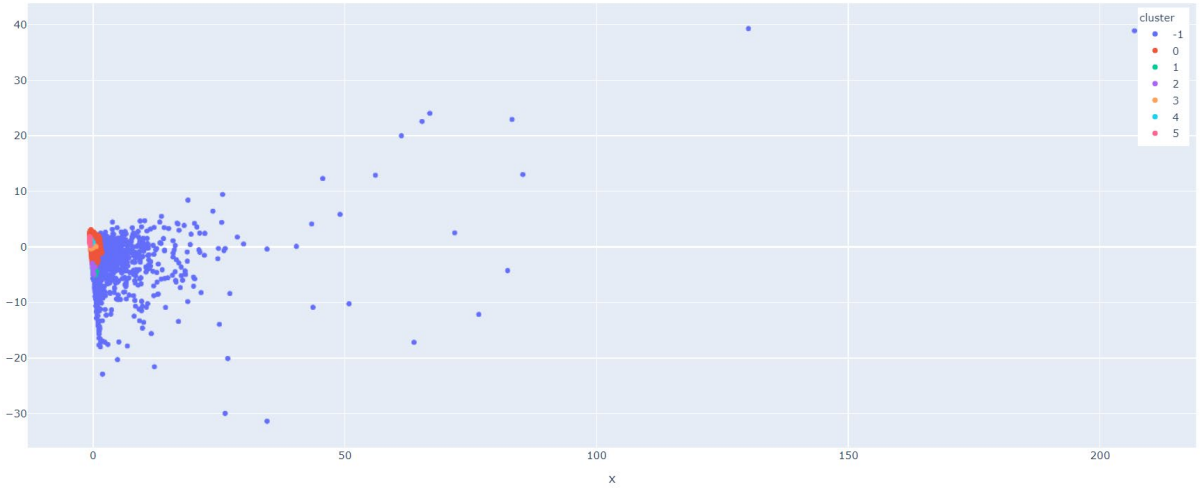
```
# dbscan
dbscan = DBSCAN(eps=2, min_samples=15).fit(data_to_fit_scaled)
```

- eps = **2** – maximálna vzdialenosť medzi 2 vzorkami na to, aby boli považované za „susedov“
- min_samples = **15** – minimálny počet vzoriek v okolí bodu, aby bol považovaný za začiatok clustra.

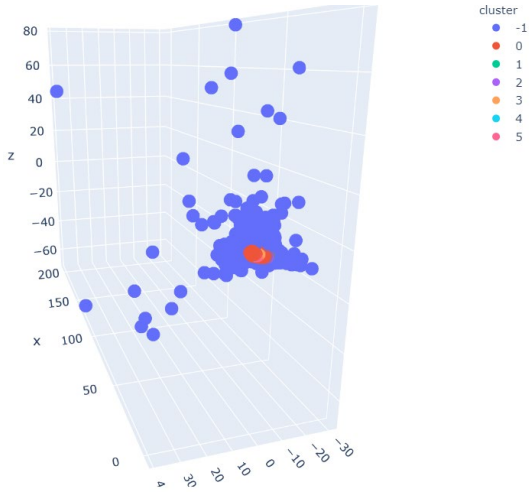
4.1 REDUKCIA DIMENZIE

Dimenziu som redukovala pomocou PCA do 2D a 3D priestoru, ktoré som vizualizovala pomocou scatter grafov. Šum je na grafoch zobrazený ako modrá časť, ktorá sa javí ako najrozptýlenejšia. Ostatné clustre sú veľmi nahusto na sebe.

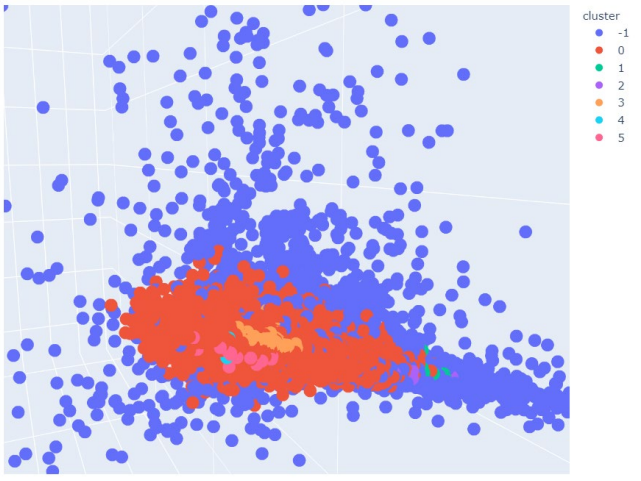
PCA – 2D



PCA – 3D

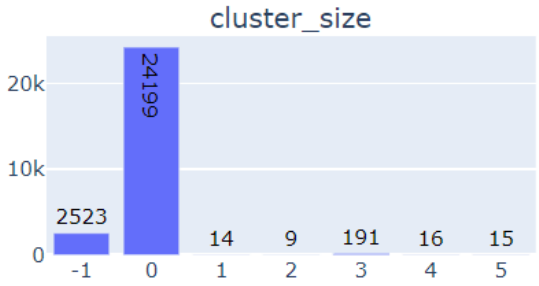


PCA – 3D (zoom)



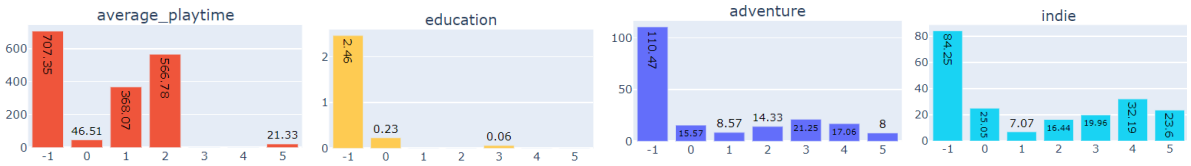
4.2 ANALÝZA

DBSCAN rozdelil vzorky veľmi nerovnomerne, v porovnaní s K-means, čo sa dá vidieť aj na PCA grafoch. Pri horeuvedených nastaveniach bolo vytvorených 5 clusterov, z ktorých cluster 0 výrazne prevláda. V clusteri -1 = šume sa nachádza okolo 2500 vzoriek.



4.2.1 Cluster -1

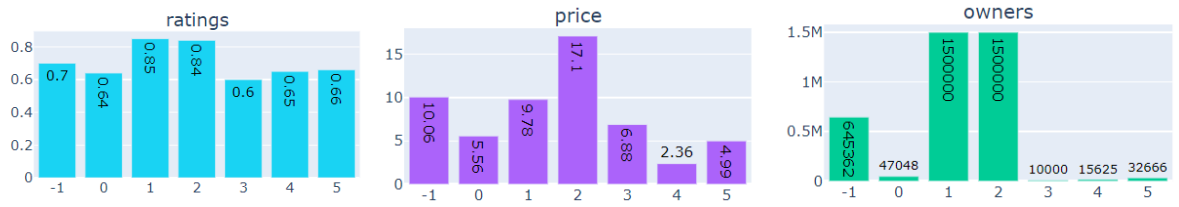
V šume sa nachádza veľa hier rôznych vlastností a kategórií, sú tu aj najhranejšie a veľmi obľúbené hry, ako aj menej známe a horšie hodnotené. To, že algoritmus zaradil do šumu veľmi veľa hier s rôznymi vlastnosťami značí o tom, že pre tento dataset je vhodnejší model s vopred určeným počtom clusterov. Zmena parametrov DBSCANu výrazne neovplyvnila nerovnomerné generovanie clusterov, pri snahe znížiť veľkosť šumu sa redukoval aj počet vytvorených clustrov.



Príklady hier z clustra -1 (šumu) – rôzne žánre, ceny, hodnotenia hier

4.2.2 Cluster 0

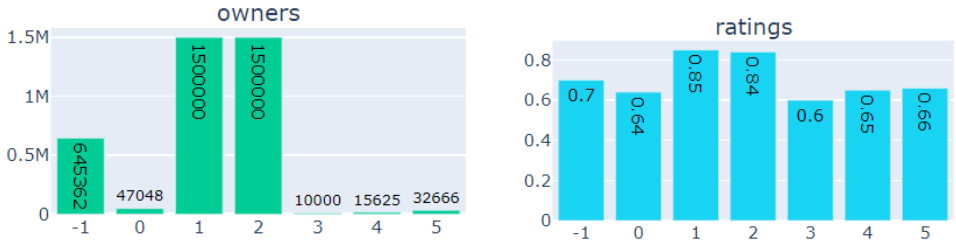
Cluster 0 je najpočetnejší, hry v tomto clustri sú na PCA grafoch zobrazené veľmi husto pri sebe. Čo sa týka žánrov, znova je to mix veľa rôznych žánrov, neprevláda ani jeden z nich. Hry tu majú relatívne nízke priemerné hodnotenie, priemerná cena je tiež pomerne nízka, čo môže byť spôsobené aj veľkosťou clusteru. Tieto hry by mohli byť označené ako menej obľúbené hry.



Príklady hier z clustra 0 – nízke hodnotenie, lacné, neznáme

4.2.3 Cluster 1 a 2

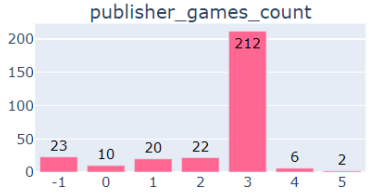
Hier v týchto clustroch je veľmi málo. Obidva clustre majú rovnaké počty vlastníkov a skoro rovnaké a vysoké priemerné hodnotenia. Označila by som hry v týchto clustroch ako populárne, s tým, že by som tieto 2 clustre spojila do jedného, kvôli podobným vlastnostiam.



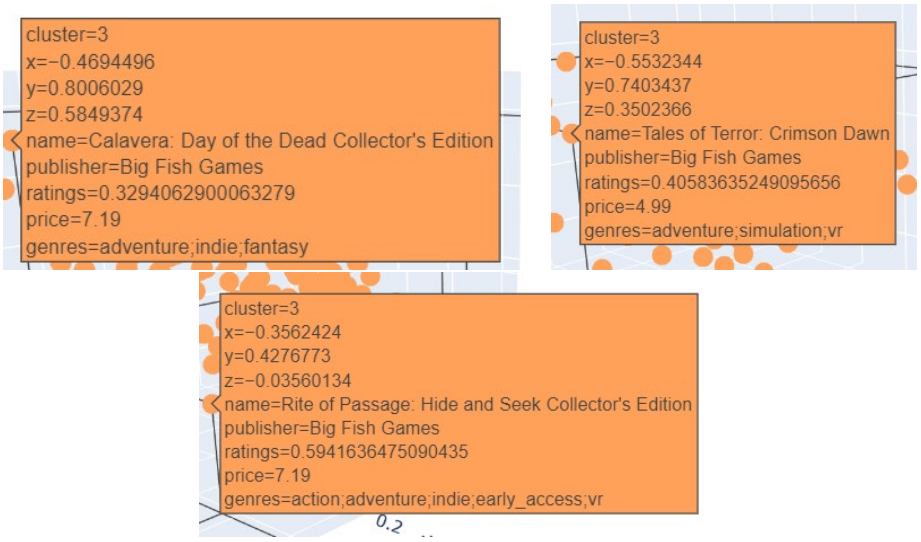
Príklady hier z clustra 1 – populárne, známe hry

4.2.4 Cluster 3

V clustri 3 bola jediná výrazná vlastnosť – počet vydaných hier publishera. Tento cluster by mohol, rovnako ako pri K-means, reprezentovať veľmi aktívnych vydavateľov hier.



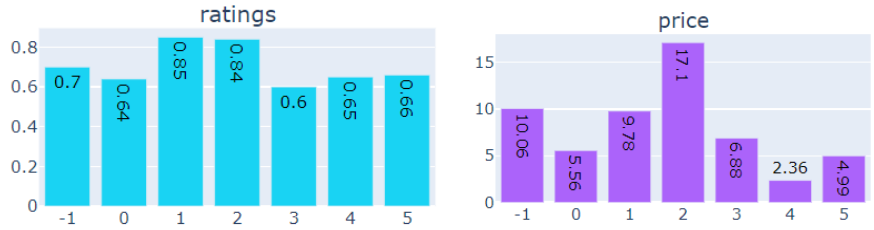
Príklady hier z clustra 1 – publisher s veľkým počtom vydaných hier



Celý cluster obsahuje hry iba od jedného publishera – Big Fish Games, ktorý je na 1. mieste v množstve vydaných hier.

4.2.5 Cluster 4 a 5

Oba clustre obsahujú veľmi málo hier a nevyniká žiadny zo žánrov. Majú nízke hodnotenia a nízku cenu.



Príklady hier z clustra 4 a 5 – nízke hodnotenia, neznáme