

KAUNO TECHNOLOGIJOS UNIVERSITETAS

INFORMATIKOS FAKULTETAS

Intelektikos pagrindai

Laboratorinis darbas Nr. 1

Duomenų apdorojimas rinkinio analizė

**Atliko:**

IFF-7/13 grup. Stud:

Tautvydas Dikšas

**Vertino:**

Lekt. Germanas Budnikas

doc. Agnė Paulauskaitė-Tarasevičienė

Turinys

[1. Užduotis 3](#_Toc37088489)

[2. Tolydinio tipo duomenų rinkinio kokybės analizė 3](#_Toc37088490)

[3. Kategorinio duomenų rinkinio kokybės analizė 6](#_Toc37088491)

[4. Atributų histogramos 8](#_Toc37088492)

[5. Duomenų kokybės problemų identifikavimas 17](#_Toc37088493)

[6. Atributų sąryšių nustatymas 17](#_Toc37088494)

[6.1. Scatter Plot Matrix diagrama: 19](#_Toc37088495)

[7. Koreliacijos matrica 20](#_Toc37088496)

[8. Išvados 20](#_Toc37088497)

# Užduotis

Pasirinkti duomenys turintys 16 stulpelių bei 16719 įrašų. Iš jų yra 10 tolydiniai ir 6 kategoriniai.

Laboratoriniui darbui atlikti naudotasi Python programavimo kalba.

Taip pat naudojamos bibliotekos: Pandas – duomenų skaitymui ir diagramų braižymui.

**Naudojamos bibliotekos kode**:



**Duomenų nuoroda:** <https://www.kaggle.com/sidtwr/videogames-sales-dataset#Video_Games_Sales_as_at_22_Dec_2016.csv>

# Tolydinio tipo duomenų rinkinio kokybės analizė

Šioje lentelėje pateikta tolydinio tipo atributų kokybės analizė. Lentelė parodo, kad yra stulpelių, kuriems nemažai trūksta duomenų, todėl potencialiai galima jas nenaudoti.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pavadinimas | Bendras reikšmių skaičius | Trūkstamų skaičius | Kardinalumas | Minimali reikšmė | Maksimali reikšmė | 1-asis kvartilis | 2-asis kvartilis | Vidurkis | Mediana | Standartinis nuokrypis |
| Year\_of\_Release | 16719 | 1.609% | 39 | 1980.0 | 2020.0 | 2003.0 | 2010.0 | 2006.487 | 2007.0 | 5.899 |
| NA\_Sales | 16719 | 0.0% | 402 | 0.0 | 41.36 | 0.0 | 0.24 | 0.263 | 0.08 | 0.855 |
| EU\_Sales | 16719 | 0.0% | 307 | 0.0 | 28.96 | 0.0 | 0.11 | 0.145 | 0.02 | 0.524 |
| JP\_Sales | 16719 | 0.0% | 244 | 0.0 | 10.22 | 0.0 | 0.04 | 0.078 | 0.0 | 0.318 |
| Other\_Sales | 16719 | 0.0% | 155 | 0.0 | 10.57 | 0.0 | 0.03 | 0.047 | 0.01 | 0.193 |
| Global\_Sales | 16719 | 0.0% | 629 | 0.01 | 82.53 | 0.06 | 0.47 | 0.534 | 0.17 | 1.637 |
| Critic\_Score | 16719 | 51.331% | 82 | 13.0 | 98.0 | 60.0 | 79.0 | 68.968 | 71.0 | 13.971 |
| Critic\_Count | 16719 | 51.331% | 106 | 3.0 | 113.0 | 12.0 | 36.0 | 26.361 | 21.0 | 18.983 |
| User\_Score | 16719 | 54.603% | 95 | 0.0 | 9.7 | 6.4 | 8.2 | 7.125 | 7.5 | 1.505 |
| User\_Count | 16719 | 54.603% | 888 | 4.0 | 10665.0 | 10.0 | 81.0 | 162.23 | 24.0 | 561.245 |

Tolydinių atributų lentelės rezultatų realizavimui naudojamas kodas:





# Kategorinio duomenų rinkinio kokybės analizė

Šioje lentelėje pateikta kategorinio tipo atributų kokybės analizė. Kategorinių atributų problema yra aukštas kardinalumas.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pavadinimas** | **Bendras reikšmių skaičius** | **Trūkstamų skaičius** | **Kardinalumas** | **Moda** | **Modos dažnumas** | **Modos dažnumas procentais** | **2-oji moda** | **2-osios modos dažnumas** | **2-osios modos dažnumas procentais** |
| Name | 16719 | 0.012% | 11562 | Need for Speed: Most Wanted | 12 | 0.072% | Ratatouille | 9 | 0.054% |
| Platform | 16719 | 0.0% | 31 | PS2 | 2161 | 12.925% | DS | 2152 | 12.872% |
| Genre | 16719 | 0.012% | 12 | Action | 3370 | 20.157% | Sports | 2348 | 14.044% |
| Publisher | 16719 | 0.323% | 582 | Electronic Arts | 1356 | 8.111% | Activision | 985 | 5.892% |
| Developer | 16719 | 39.614% | 1696 | Ubisoft | 204 | 1.22% | EA Sports | 172 | 1.029% |
| Rating | 16719 | 40.487% | 8 | E | 3991 | 23.871% | T | 2961 | 17.71% |

Kategorinių atributų lentelės rezultatų realizavimui naudojamas kodas:



# Atributų histogramos

Tolydinių atributų histogramoms realizuoti naudojamas kodas:

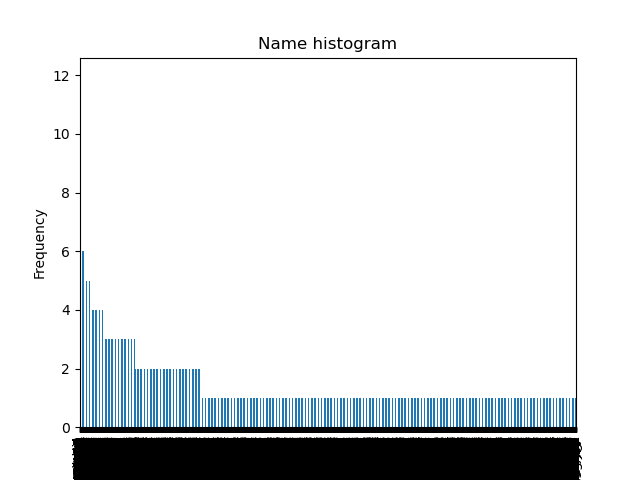


Kategorinio tipo atributų histogramoms realizuoti naudojamas kodas:



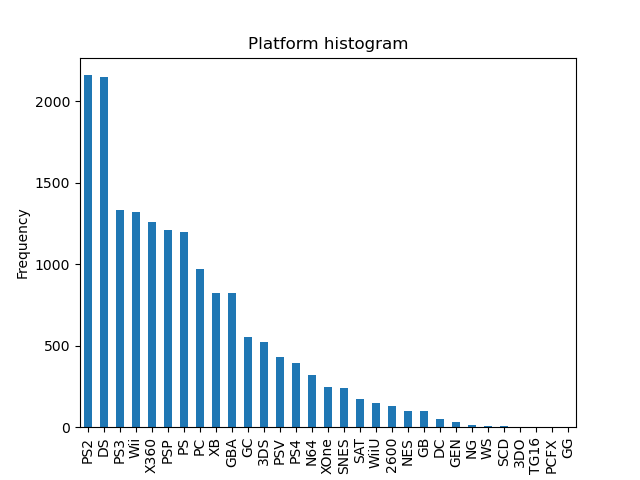
**Histogramos:**

Pav. 4.1 matome kaip atrodo situaciją, kai yra daug skirtingų pavadinimų t.y. didelis kardinalumas.



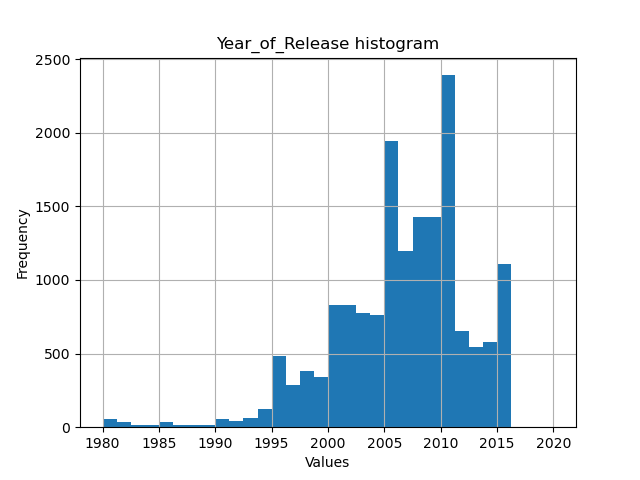
pav. 4.1 Atributo Name histograma

Pav. 4.2 mažėjančiai pasiskirsčiusias platformų kategorijas. ,,Right-skewed” pasiskirstymas.



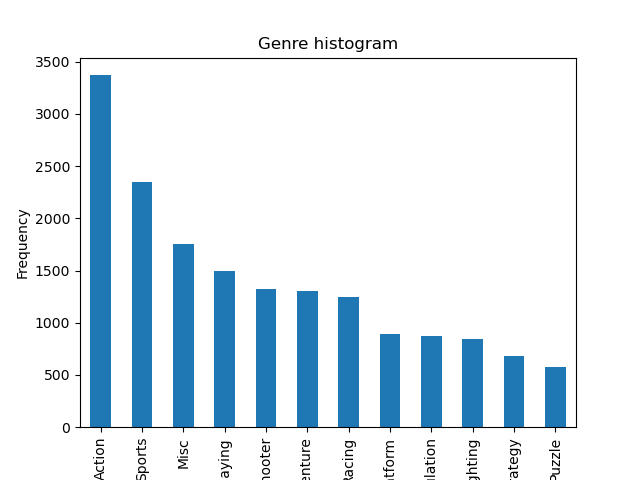
pav. 4.2 Atributo Platform histograma

Pav. 4.3 atvaizduoja year\_of\_Release histogramą. ,,Random” pasiskirstymas.



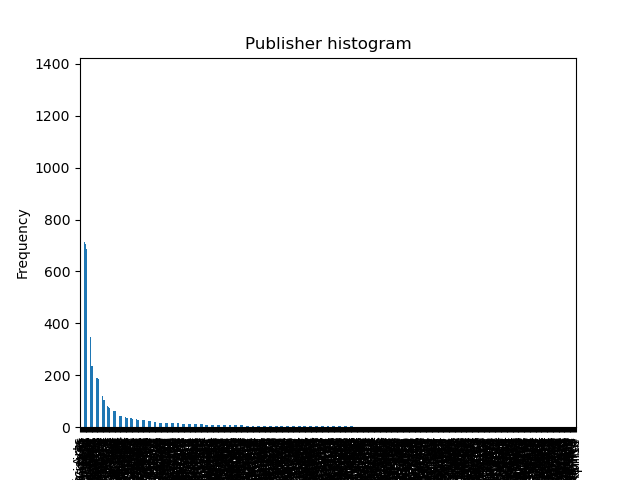
pav. 4.3 Atributo Year of Release histograma

Pav. 4.4 atvaizduoja Genre histogramą. Dėšinėje nukreiptas pasiskirstymas. ,,Right-skewed” pasiskirstymas



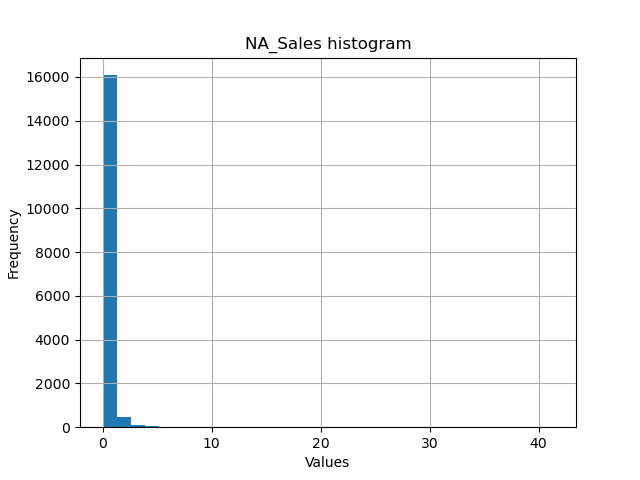
pav. 4.4 Atributo Genre histograma

Pav. 4.5 atvaizduoja publisher histogramą. Diagramoje matoma didelio kardinalumo problema. Eksponentinis mažėjantis pasiskirstymas.



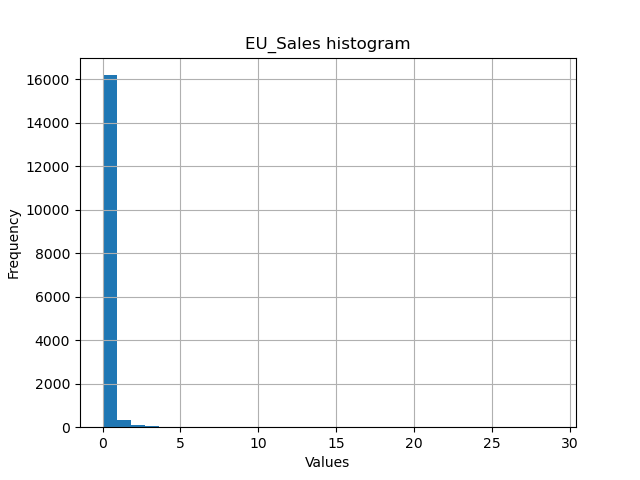
pav. 4.5 Atributo Publisher histograma

Pav. 4.6 atvaizduoja NA\_Sales histogramą. Joje matome, kad turi ekstremalių reikšmių, todėl vienas ,,bin” yra žymiai aukštesnis už kitus. Reikia pasirinkti interval, kad diagrama išsilygintų.



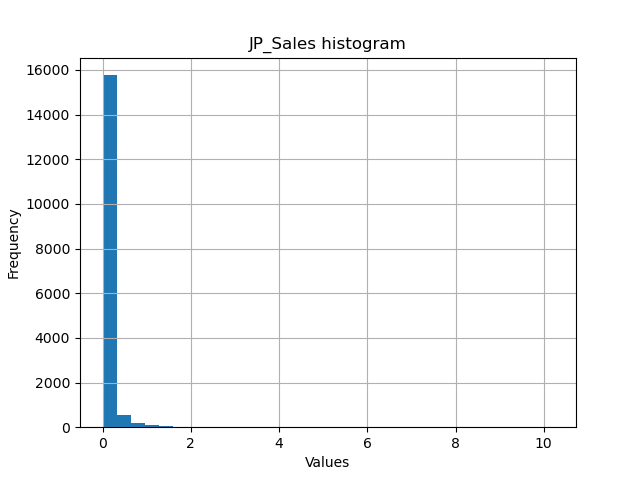
pav. 4.6 Atributo NA Sales histograma

Pav. 4.7 atvaizduoja EU\_Sales histogramą. Joje matome, kad turi ekstremalių reikšmių, todėl vienas ,,bin” yra žymiai aukštesnis už kitus. Reikia pasirinkti interval, kad diagrama išsilygintų.



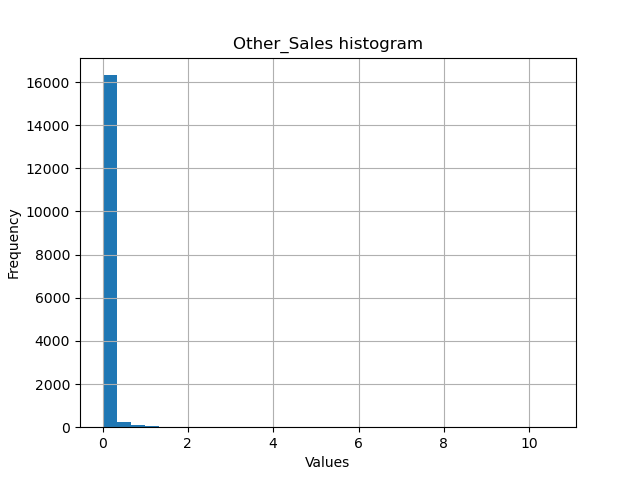
pav. 4.7 Atributo EU Sales histograma

Pav. 4.8 atvaizduoja JP\_Sales histogramą. Joje matome, kad turi ekstremalių reikšmių, todėl vienas ,,bin” yra žymiai aukštesnis už kitus. Reikia pasirinkti interval, kad diagrama išsilygintų.



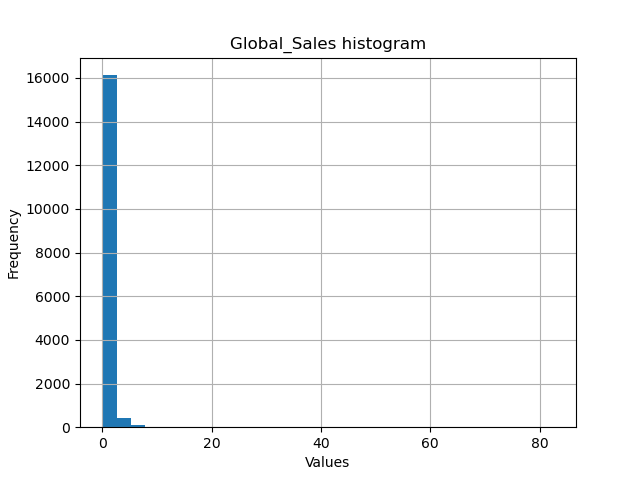
pav. 4.8 Atributo JP Sales histograma

Pav. 4.9 atvaizduoja Other\_Sales histogramą. Joje matome, kad turi ekstremalių reikšmių, todėl vienas ,,bin” yra žymiai aukštesnis už kitus. Reikia pasirinkti interval, kad diagrama išsilygintų.



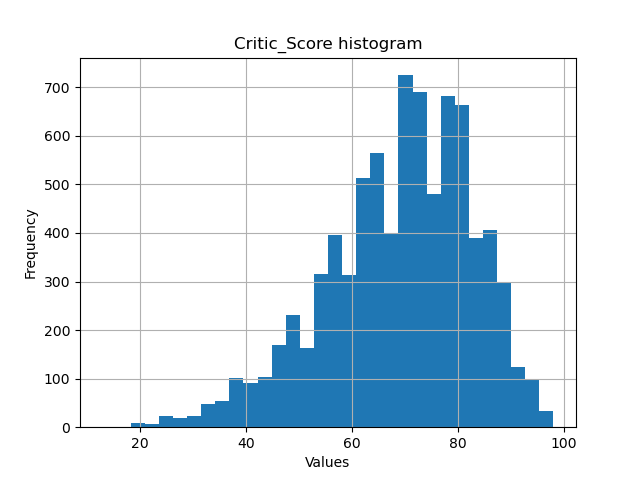
pav. 4.9 Atributo Other Sales histograma

Pav. 4.10 atvaizduoja Global\_Sales histogramą. Joje matome, kad turi ekstremalių reikšmių, todėl vienas ,,bin” yra žymiai aukštesnis už kitus. Reikia pasirinkti interval, kad diagrama išsilygintų.



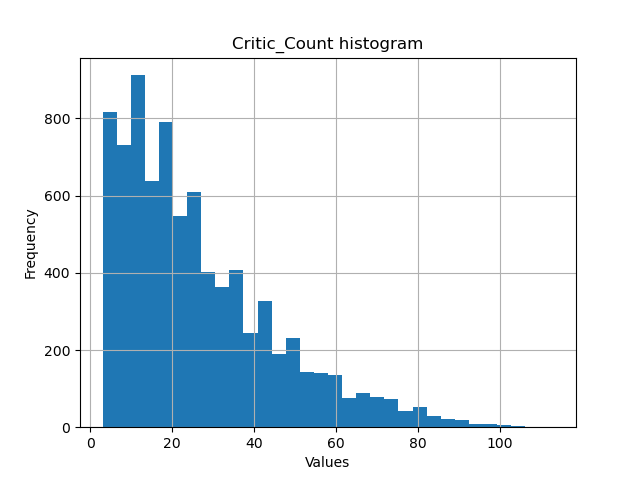
pav. 4.10 Atributo Global Sales histograma

Pav. 4.11 parodo atributo Critic\_Score histogramą. Beveik normalusis pasiskirstymas



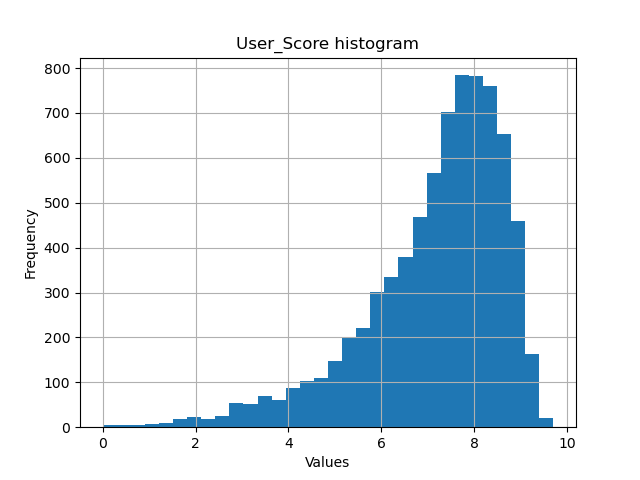
pav. 4.11 Atributo Critic Score histograma

Pav. 4.12 atvaizduoja Critic\_Count histogramą. ,,Right-Skewed” pasiskirstymas



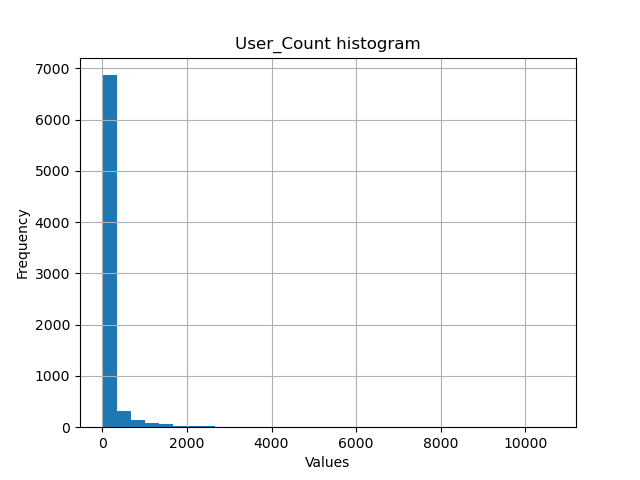
pav. 4.12 Atributo Critic Count histograma

Pav. 4.13 atvaizduota User\_Score histograma. Eksponentinis didėjantis pasiskirstymas.



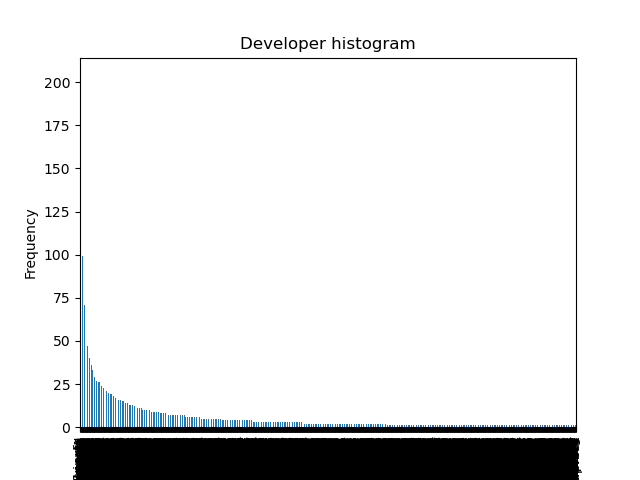
pav. 4.13 Atributo User Score histograma

Pav. 4.14 atvaizduoja User\_Count histogramą. Atributo duomenyse yra ekstremalios reikšmės, todėl reikia pasirinkti intervalą, kad diagrama susilygintų.



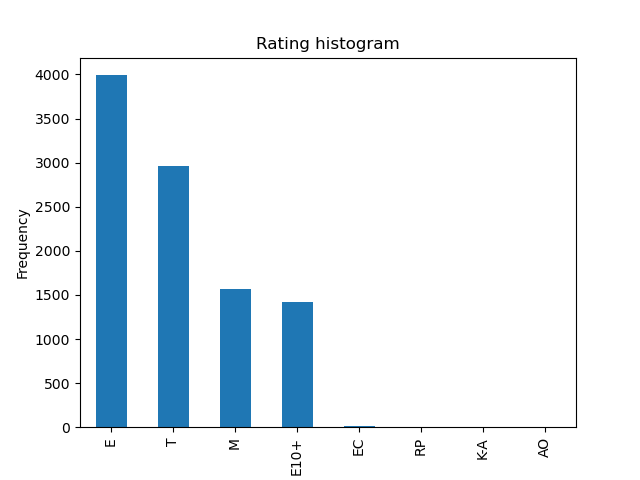
pav. 4.14 Atributo User Count histograma

Pav. 4.15 parodo atributo Developer histrogramą. Eksponentinis mažėjantis



pav. 4.15 Atributo Developer histograma

Pav. 4.16 atvaizduoja Rating atributo diagrama. Dešinėje pusėje nukreiptas pasiskirstymas (mažėjantis) t.y. ,,right skewed”



pav. 4.16 Atributo Rating histograma

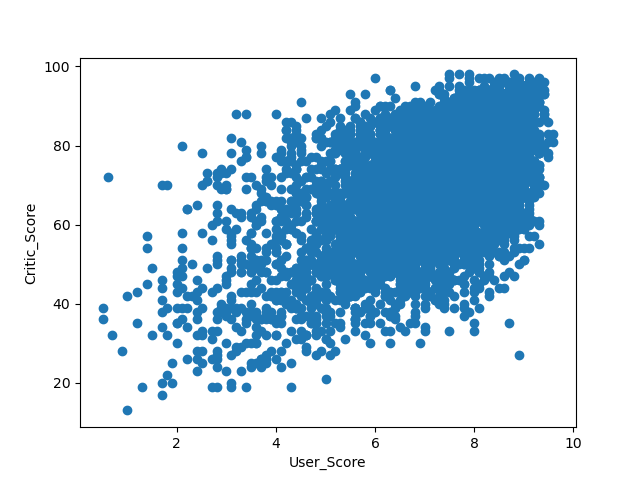
# Duomenų kokybės problemų identifikavimas

Duomenų rinkinyje buvo rastos kelios kokybės problemos. Pirma problema buvo trūkstamos reikšmės. Atvaizduojant histogramas tos eilutės buvo praleidžiamos. Antroji problema – ekstremalios reikšmės. Šios problemos sprendimo būdas gali būti intervalo nustatymas.

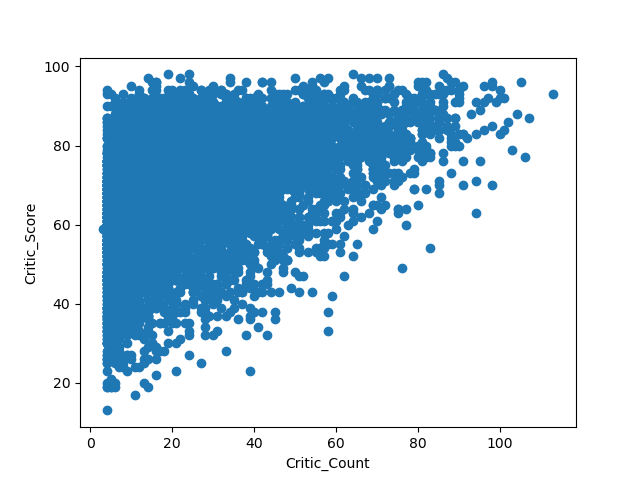
# Atributų sąryšių nustatymas

Stipri koreliacija:

Stipri koreliacija parodo, kad yra ryšys tarp dviejų atributų. Būtent tai parodo Pav. 6.1 ir Pav. 6.2



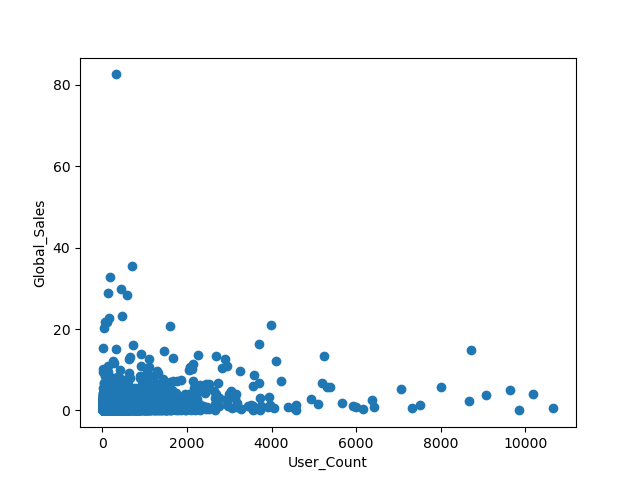
pav. 6.1 Atributų Critic\_Score ir User\_Score priklausomybė



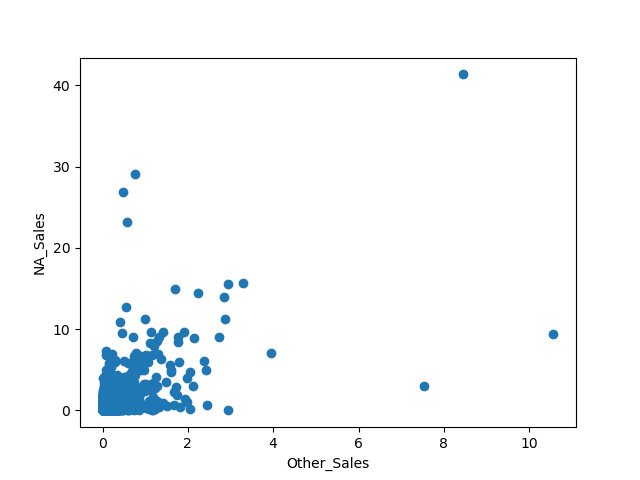
pav. 6.2 Atributų Critic\_Score ir Critic\_Count priklausomybė

Silpna koreliacija:

Silpna koreliacija parodo, kad ryšio tarp atributų tikriausiai nėra. Pav. 6.3 ir Pav. 6.4 parodo, kad pas juos silpna koreliacija ir pasikeitus duomenims jų diagramai įtakos neturės.



pav. 6.3 Atributų Global\_Sales ir User\_Count priklausomybė



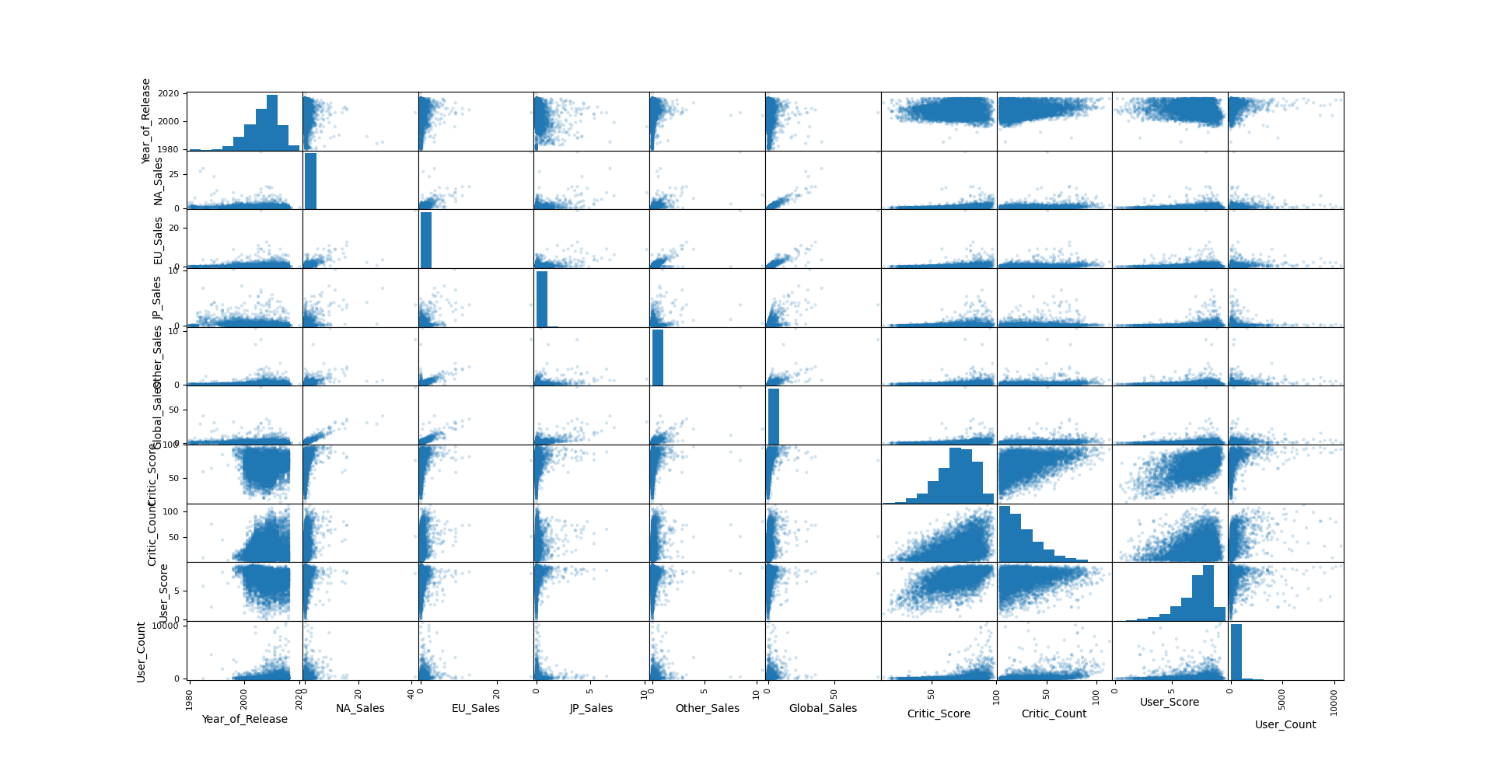
pav. 6.4 Atributų NA\_Sales ir Other\_Sales priklausomybė

Realizavimui naudotas kodas:



## Scatter Plot Matrix diagrama:

Pav. 6.1 atvaiduoja scatter plot matrix diagramą, kuri parodo atributų priklausomybes. Diagramoje esančios histogramos rodo, kad priklausomybė lyginima su savimi.



pav. 6.5 SPLOM diagrama

Diagramos atvaizdavimas kodu:

def splom(data):

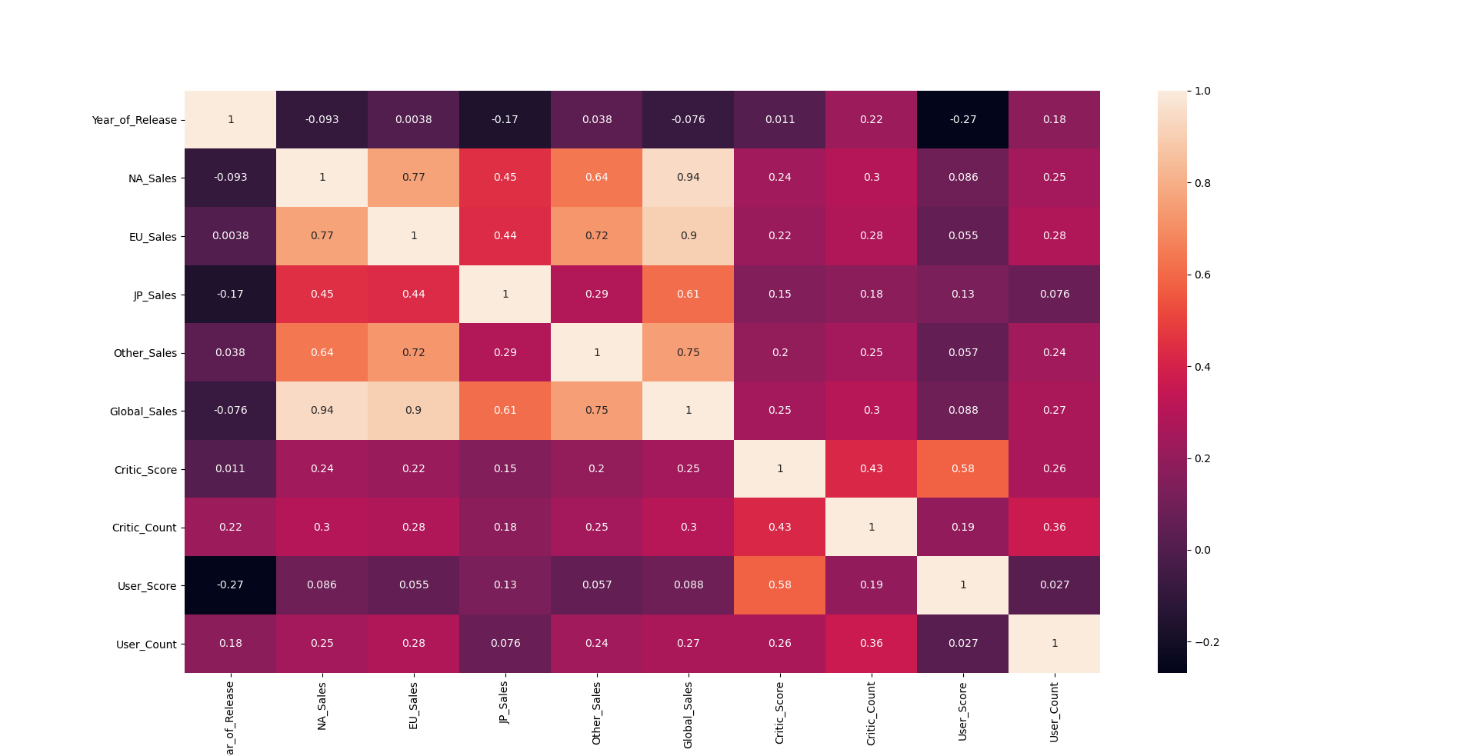
    cont, categ = sepheaders(data)

    pd.plotting.scatter\_matrix(data, alpha=0.2)

    plt.show()

# Koreliacijos matrica

Pav. 7.1 išspausdinta tolydinių atributų koreliacijos matrica. Joje matome, kad stipri koreliacija yra tarp NA\_Sales ir Global\_Sales, NA\_Sales ir EU\_Sales, EU\_Sales ir Global\_Sales. Nėra koreliacijos matoma su Year\_of\_Release atributu, todėl sprendimas būtų išmesti jį arba paversti kategoriniu.



pav. 7.1 Tolydinių atributų koreliacijos matrica

Šios diagramos atvaizdavimas kodu:



# Išvados

* Analizuojant duomenų rinkinį nustatyta, kurie atributai tarpusavyje priklausomi
* Atliekant duomenų rinkinio analizę pastebėta, kad yra stulpelių, kuriems trūksta >50% duomenų
* Duomenų rinkinyje pastebėta daug stulpelių, kurie visiškai nepriklausantys, todėl buvo galima juos ištrinti.