Paveikslėlis, kuriame yra juodas, tamsa

Automatiškai sugeneruotas aprašymas

**Kauno technologijos universitetas**

Informatikos fakultetas

**P176B101 Intelektikos pagrindai**

**Laboratorinis darbas Nr. 1**

|  |  |
| --- | --- |
|  |  |
| **Nedas Liaudanskis IFF-1/9**  Studentas | (parašas) (data) |
|  |  |
| **dėst. Nečiūnas Audrius**  Dėstytojas | (parašas) (data) |
|  |  |

**KAUNAS, 2024**

Turinys

[1. Įvadas 3](#_Toc159493528)

[2. Duomenų rinkinys 4](#_Toc159493529)

[3. Duomenų rinkinio kokybės analizė 5](#_Toc159493530)

[ Tolydinis tipas 5](#_Toc159493531)

[ Kategorinis tipas 6](#_Toc159493532)

[4. Duomenų kokybės problemų identifikavimas ir sprendimas 8](#_Toc159493533)

[5. Atributų grafikai 9](#_Toc159493534)

[ Tolydinio tipo histogramos 9](#_Toc159493535)

[ Kategorinio tipo stulpelinės diagramos 11](#_Toc159493536)

[ „Scatter plot“ ir SPLOM diagrama 13](#_Toc159493537)

[6. Kovariacija ir koreliacija 14](#_Toc159493538)

[ Kovariacija 14](#_Toc159493539)

[ Koreliacija 15](#_Toc159493540)

[7. Duomenų normalizacija 16](#_Toc159493541)

[8. Išvados 17](#_Toc159493542)

1. Įvadas

Laboratorinio darbo tikslas yra išsirinkti duomenų rinkinį, jį apdoroti, sutvarkyti ir sutvarkius išanalizuoti duomenis. Darbo eiga:

1. Pasirinkti duomenų rinkinį (Rinkinyje turi būti nemažiau kaip 500 atributų, n ir nemažiau kaip 8 stulpeliai).
2. Atlikti duomenų rikinio kokybės analizę.
3. Kiekvienam tolydinio tipo atributui paskaičiuoti:

* Bendrą reikšmių skaičių.
* Trūkstamų reikšmių procentą.
* Kardinalumą.
* Minimumą ir maksimumą.
* 1-ąją ir 3-ąją kvartilius.
* Vidurkį.
* Medianą.
* Standartinį nuokrypį.

1. Kiekvienam kategorinio tipo atributui paskaičiuoti:

* Bendrą reikšmių skaičių.
* Trūkstamų reikšmių procentą.
* Kardinalumą.
* Modą.
* Modos dažnumą.
* Modos procentinę reikšmę.
* 2-ąją modą.
* 2-osios modos dažnumo reikšmę.
* 2-osios modos procentinę reikšmę.

1. Identifikuoti duomenų kokybės problemas: trūkstamas reikšmes, kardinalumo problemas, triukšmus– ekstremalias reikšmes.
2. Nubraižyti atributų histogramas.
3. Sąryšius tarp atributų atvaizduoti vizualiai.
4. Paskaičiuoti kovariacijos ir koreliacijos reikšmes tarp tolydinio tipo atributų ir grafiškai atvaizduoti koreliacijos matricą.
5. Atlikti duomenų normalizaciją.
6. Duomenų rinkinys

Norint atlikti laboratorinį darbą, reikia pasirinkti duomenų rinkinį, kuriame būtų nemažiau kaip 500 įrašų ir jis tūrėtų nemažiau kaip 8 stulpelius (atributus). Atributus turi sudaryti bent vienas, tolydinio tipo atributas ir bent vienas kategorini tipo atributas. Rinkinių pasirinkimai, buvo pateikti 3-ose svetainėse:

* <https://www.kaggle.com/datasets>
* <https://archive.ics.uci.edu/datasets.php>
* <https://vincentarelbundock.github.io/Rdatasets/datasets.html>

Laboratoriniam darbui buvo pasirinktas duomenų rinkinys pavadinimu: Video Game Sales with Ratings.

<https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings?select=Video_Games_Sales_as_at_22_Dec_2016.csv>

Duomenų rinkinį sudaro, 16 stulpelių (atributų) ir 16719 įrašai. Jame yra 6 kategoriniai atributai ir 10 tolydiniai atributai.

Duomenų rinkinio atributai:

* Name – Žaidimo pavadinimas.
* Platform- Platforma kurioje buvo išleistas žaidimas.
* Year\_of\_Release – metai, kuriais buvo išleistas žaidimas.
* Genre – žaidimo žanras.
* Publisher – Firma, kuri išleido šį žaidimą.
* NA\_Sales – Žaidimo pardavimai Šiaurės Amerikoje (milijonais).
* EU\_Sales - Žaidimo pardavimai Europoje (milijonais).
* JP\_Sales - Žaidimo pardavimai Japonijoje (milijonais).
* Other\_Sales - Žaidimo pardavimai kituose žemynuose (milijonais).
* Global\_Sales – totalus parduotų kopijų skaičius visame pasaulyje (milijonais).
* Critic\_Score – Kritikų vertinimas (nuo 0 iki 100 taškų).
* Critic\_Count – Kritikų skaičius.
* User\_Score – Naudotojų vertinimas (nuo 0 iki 100 taškų).
* User\_Count – Naudotojų skaičius.
* Developer – Firma, kuri sukūrė žaidimą.
* Rating – ESRB vertinimas (<https://www.esrb.org/ratings-guide/>)

1. Duomenų rinkinio kokybės analizė

Duomenys, dažnai turi įvairiausių defektų, kurie gali mums duoti netinkamus rezultatus. Todėl pirmiausia reikia atlikti duomenų rinkinio kokybinę analizę. Šios analizės pagalba galėsime surasti defektus, kurių sprendimą pateiksime sekančiame etape.

* Tolydinis tipas

Kiekvienam atributui reikia apskaičiuoti:

* Bendrą reikšmių skaičių.
* Trūkstamų reikšmių procentą.
* Kardinalumą.
* Minimumą ir maksimumą.
* 1-ąją ir 3-ąją kvartilius.
* Vidurkį.
* Medianą.
* Standartinį nuokrypį.

Tolydinio tipo duomenų rinkinio kokybės analizės rezultatai matomi 1 lentelėje.

Iš gautų rezultatų galime pastebėti, jog yra keletas reikšmių, kurių trūkstamų reikšmių skaičius artėja prie 50% (“Critic\_Score”, “Critic\_Count”, “User\_Score”, “User\_Count”), tai labai didelis trūkstamų reikšmių skaičius. Dėl didelių trūkstamų reikšmių šiuos stulpelius reikės pašalinti, nes jie kelia riziką tikslingam skaičiavimam tolimesniuose etapuose. Nepaminėtuose atributuose, trūkstamų reikšmių skaičius beveik, nepasireiškia. Išskyrus atribute „Year\_Of\_Release“, kuriame trūksta labai nedaug duomenų, 1,6%. Įrašus, kuriose trūksta atributo „Year\_Of\_Release“, teks pašalinti. Kadangi tai nedidelis įrašų skaičius, tai neturės didelės įtakos galutiniams skaičiavimams.

Lentelėje, taip pat galime pamatyti, jog daugumoje atributų („NA\_Sales“, „JP\_Sales“, „EU\_Sales“, „Other\_Sales“, „Global\_Sales“, „User\_Score“), 1-ojo kvartilio reikšmės yra lygios 0. Tai pasako, jog 25% duomenų yra lygūs arba labai artimi 0. Tai gali būti dėl t, jog pardavimų reikšmės yra rašomos milijonais, mažesni žaidimų pardavimai, turės daug mažesnes reikšmes, jeigu skaičiuosime milijonais. Todėl dauguma pardavimų artėja prie 0.

*Lentelė 1, Tolydini tipo atributų skaičiavimų reikšmės,, duomenų rinkinio analizei.*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atributo pavadinimas** | **Kiekis** | **Trūkstamos reikšmės, %** | **Kardinalumas** | **Minimali reikšmė** | **Maksimali reikšmė** | **1-asis kvartilis** | **3-iasis kvartilis** | **Vidurkis** | **Mediana** | **Standartinis nuokrypis** |
| Year\_of\_Release | 16719 | 1,6089 | 39 | 1980 | 2020 | 2003 | 2010 | 2006 | 2007 | 5,879 |
| NA\_Sales | 16719 | 0,0000 | 402 | 0 | 41,36 | 0 | 0,24 | 0,263 | 0,08 | 0,8135 |
| EU\_Sales | 16719 | 0,0000 | 307 | 0 | 28,96 | 0 | 0,11 | 0,145 | 0,02 | 0,5033 |
| JP\_Sales | 16719 | 0,0000 | 244 | 0 | 10,22 | 0 | 0,04 | 0,078 | 0 | 0,3088 |
| Other\_Sales | 16719 | 0,0000 | 155 | 0 | 10,57 | 0 | 0,03 | 0,047 | 0,01 | 0,1867 |
| Global\_Sales | 16719 | 0,0000 | 629 | 0,01 | 82,53 | 0 | 0,47 | 0,534 | 0,17 | 1,5479 |
| Critic\_Score | 16719 | 51,3308 | 82 | 13 | 98 | 60 | 79 | 68,97 | 71 | 13,9382 |
| Critic\_Count | 16719 | 51,3308 | 106 | 3 | 113 | 12 | 36 | 26,36 | 21 | 18,9805 |
| User\_Score | 16719 | 40,0981 | 96 | 0 | 9,7 | 6,4 | 8,2 | 7,125 | 7,5 | 1,5 |
| User\_Count | 16719 | 54,6025 | 888 | 4 | 10665 | 10 | 81 | 162,2 | 24 | 561,2823 |

* Kategorinis tipas

Kiekvienam atributui reikia apskaičiuoti:

* Bendrą reikšmių skaičių.
* Trūkstamų reikšmių procentą.
* Kardinalumą.
* Modą.
* Modos dažnumą.
* Modos procentinę reikšmę.
* 2-ąją modą.
* 2-osios modos dažnumo reikšmę.
* 2-osios modos procentinę reikšmę.

Kategorinio tipo duomenų rinkinio kokybės analizės rezultatai matomi 2 lentelėje.

Iš gautų rezultatų galime pastebėti, jog yra keletas reikšmių, kurių trūkstamų reikšmių skaičius artėja prie 50% (“Developer”, “Rating”). Tai panašiai kaip ir su tolydinio tipo duomenimis, šiuos stulpelius reikės tvarkyti, nes jų nemodifikavus gausime neteisingus atsakymus sekančiuose etapuose. Likę atributai, turi nedaug trūkstamų reikšmių, tačiau ir jas teks tvarkyti, greičiausiai pašalinti įrašus, kurie neturi tam tikro atributo reikšmės.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Atributo pavadinimas** | **Platform** | **Genre** | **Publisher** | **Developer** | **Rating** | **Name** |
| **Kiekis** | 16719 | 16719 | 16719 | 16719 | 16719 | 16719 |
| **Trūkstamos reikšmės, %** | 0**%** | 0,0120**%** | 0,3230**%** | 39,6136**%** | 40,4869**%** | 0,0120**%** |
| **Kardinalumas** | 31 | 12 | 581 | 1696 | 8 | 11562 |
| **Moda** | PS2 | Action | Electronic Arts | Ubisoft | E | Need for Speed: Most Wanted |
| **Modos dažnumas** | 2161 | 3370 | 1356 | 204 | 3991 | 12 |
| **Moda, %** | 12,93% | 20,16% | 8,14% | 2,02% | 40,11% | 0,07% |
| **2-oji Moda** | DS | Sports | Activision | EA Sports | T | Ratatoille |
| **2-osios modos dažnumas** | 2152 | 2348 | 985 | 172 | 2961 | 9 |
| **2-oji moda, %** | 12,87% | 14,05% | 5,91% | 1,70% | 29,76% | 0,05% |
|  |  |  |  |  |  |  |

*Lentelė 2, parodo kategorinio tipo atributų skaičiavimus, duomenų rinkinio analizei.*

1. Duomenų kokybės problemų identifikavimas ir sprendimas

Pagrindinės problemos, kurias galime pastebėti iš abiejų lentelių, yra tai, jog yra nemažai atributų su trūkstamomis reikšmėmis. Trūkstamos reikšmės sukelia daug problemų, kai mes norime atlikti tolesnius skaičiavimus. Jos iškreipia atsakymus ir histogramas. Todėl duomenis reikia sutvarkyti. Atributus suskirstysime į dvi kategorijas:

* Atributai, kurie artėja prie 50% trūkstamų reikšmių (“Critic\_Score”, “Critic\_Count”, “User\_Score”, “User\_Count”, “Developer”, “Rating”).

.

* Atributai, kurie turi nedaug trūkstamų reikšmių („Year\_of\_Release“, „Genre“, „Publisher“, „Name“).

Atributai, kurie artėja prie 50%, turi per daug trūkstamų reikšmių, todėl juos reikės pašalinti iš tolimesnių skaičiavimų. Šaliname šiuos stulpelius(atributus) ir duomenų rinkinio. O atributus, kurie turi nedideli trūkstamų reikšmių skaičius, tvarkysime pašalinant visus įrašus, kuriuose bent vienas atributas neturi reikšmės. Atlikus šiuos pakeitimus gauname sutvarkytus duomenis lentelėje 3, 4.

*Lentelė 3, parodo kategorinio tipo atributų skaičiavimus, duomenų rinkinio analizei po modifikavimo.*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atributo pavadinimas** | **Kiekis** | **Trūkstams reikšmės** | **Kardinalumas** | **Minimali reikšmė** | **Maksimali reikšmė** | **1-asis kvartilis** | **3-iasis kvartilis** | **Vidurkis** | **Mediana** | **Standartinis nuokrypis** |
| Year\_of\_Release | 16416 | 0 | 39 | 1980 | 2020 | 2003 | 2010 | 2006,4899 | 2007 | 5,8811 |
| NA\_Sales | 16416 | 0 | 401 | 0 | 41,36 | 0 | 0,24 | 0,2641 | 0,08 | 0,8190 |
| EU\_Sales | 16416 | 0 | 307 | 0 | 28,96 | 0 | 0,11 | 0,146 | 0,02 | 0,5071 |
| JP\_Sales | 16416 | 0 | 244 | 0 | 10,22 | 0 | 0,04 | 0,0786 | 0 | 0,3113 |
| Other\_Sales | 16416 | 0 | 155 | 0 | 10,57 | 0 | 0,03 | 0,0477 | 0,01 | 0,1882 |
| Global\_Sales | 16416 | 0 | 628 | 0,01 | 82,53 | 0,06 | 0,47 | 0,5367 | 0,17 | 1,5599 |

*Lentelė 4, parodo kategorinio tipo atributų skaičiavimus, duomenų rinkinio analizei po modifikavimo.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Atributo pavadinimas** | **Name** | **Platform** | **Genre** | **Publisher** |
| **Kiekis** | 16416 | 16416 | 16416 | 16416 |
| **Trūkstamos reikšmės** | 0 | 0 | 0 | 0 |
| **Kardinalumas** | 11397 | 31 | 12 | 579 |
| **Moda** | Need for Speed: Most Wanted | PS2 | Action | Electronic Arts |
| **Modos dažnumas** | 12 | 2127 | 3307 | 1344 |
| **Moda, %** | 0,073% | 12,957% | 20,145% | 8,187% |
| **2-oji Moda** | Ratatoille | DS | Sports | Activision |
| **2-osios modos dažnumas** | 9 | 2122 | 2306 | 976 |
| **2-oji moda, %** | 0,055% | 12,926% | 14,047% | 5,945% |

1. Atributų grafikai

* Tolydinio tipo histogramos

Tolydinio tipo atributams atvaizduoti buvo naudojamos histogramos. Histogramų stulpelių skaičius buvo nustatytas naudojant formule: , kur n yra duomenų rinkinio įrašų kiekis. Panaudojus formulę gaunamas skaičius: 14. Tai pasako, jog kiekvienoje histogramoje, turės būti 14 stulpelių. Histogramos: pav. 1, 2, 3, 4, 5, 6.

Paveikslėlis, kuriame yra diagrama, Grafikas, tekstas, linija

Automatiškai sugeneruotas aprašymas

pav. 1 Histograma, parodanti atributo „Year\_of\_Release“ pasiskirstymą.

Paveikslėlis, kuriame yra tekstas, diagrama, Grafikas, linija

Automatiškai sugeneruotas aprašymas

pav. 2 Histograma, parodanti atributo „NA\_Sales“ pasiskirstymą.

Paveikslėlis, kuriame yra tekstas, ekrano kopija, linija, Grafikas

Automatiškai sugeneruotas aprašymas

pav. 3 Histograma, parodanti atributo „EU\_Sales“ pasiskirstymą.

Paveikslėlis, kuriame yra tekstas, diagrama, ekrano kopija, Grafikas

Automatiškai sugeneruotas aprašymas

pav. 4 Histograma, parodanti atributo „JP\_Sales“ pasiskirstymą.

Paveikslėlis, kuriame yra tekstas, diagrama, linija, Grafikas

Automatiškai sugeneruotas aprašymas

pav. 5 Histograma, parodanti atributo „Other\_Sales“ pasiskirstymą.

Paveikslėlis, kuriame yra diagrama, Grafikas, tekstas, linija

Automatiškai sugeneruotas aprašymas

pav. 6 Histograma, parodanti atributo „Global\_Sales“ pasiskirstymą.

* Kategorinio tipo stulpelinės diagramos

Kategorinio tipo atvaizdavimui naudojamos stulpelinės diagramos. Diagramose, kur yra per didelis kardinalumas yra rodoma tik 20 dažniausiai pasikartojančių įrašų. Diagramos: pav. 7, 8, 9, 10

Paveikslėlis, kuriame yra tekstas, ekrano kopija, diagrama, linija

Automatiškai sugeneruotas aprašymas

pav. 7 Stulpelinė diagrama rodanti atributo „Names“ 20 labiausiai pasikartojančius įrašus.

Paveikslėlis, kuriame yra tekstas, Grafikas, linija, ekrano kopija

Automatiškai sugeneruotas aprašymas

pav. 8 Stulpelinė diagrama rodanti atributo „Platform“ labiausiai pasikartojančius įrašus.

Paveikslėlis, kuriame yra tekstas, ekrano kopija, diagrama, Grafikas

Automatiškai sugeneruotas aprašymas

pav. 9 Stulpelinė diagrama rodanti atributo „Genre“ labiausiai pasikartojančius įrašus.

Paveikslėlis, kuriame yra tekstas, ekrano kopija, diagrama, linija

Automatiškai sugeneruotas aprašymas

pav. 10 Stulpelinė diagrama rodanti atributo „Publisher“ 20 labiausiai pasikartojančius įrašus.

* „Scatter plot“ ir SPLOM diagrama

"Scatter plot" ir SPLOM (santrumpa nuo "Scatter Plot Matrix") yra duomenų vizualizacijos metodai, kurie naudojami įvertinti ryšį tarp dviejų ar daugiau atributų. pav. 11.

Galime pastebėti, jog atributai ryšių vienas su kitu kaip ir neturi, nebent kalbame apie „Global\_Sales“ atributą. Šio atributo „Scatter plot“ dergiamose galime matyti, jog su kitais pardavimais šis atributas turi ryšį ir jeigu vienas atributas kyla tai kils ir „Global\_Sales“ atributo reikšmė.

Paveikslėlis, kuriame yra tekstas, ekrano kopija, diagrama, linija

Automatiškai sugeneruotas aprašymas

pav. 11 Atvaizduoja „Scatter plot“ ir SPLOM gautas diagramas.

1. Kovariacija ir koreliacija

* Kovariacija

Kovariacija parodo dviejų atributų ryšio stiprumą vienas tarp kito. Kovariacija apskaičiuojama naudojant formule:

Čia a – pirmasis atributas, b – antrasis atributas, 𝑎x – pirmojo atributo vidurkis, 𝑏x – antrojo atributo vidurkis, n – duomenų rinkinio įrašų kiekis.

Skaičiavimai naudojant formulę pateikti lentelėje 5.

Pagal gautus atsakymus galime pastebėti, jog tie laukai kurie yra teigiami, turi tam tikrą ryšį vienas nuo kito. Kuo didesnis skaičius tuo stipresnis ryšys. Didžiausią ryšį turi „NA\_Sales“ su Global\_Sales. Didžiausio ryšio skaitinė vertė: 1,2025.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year\_of\_Release |  |  |  |  |  |  |
| NA\_Sales | -0,4448 |  |  |  |  |  |
| EU\_Sales | 0,01170 | 0,3184 |  |  |  |  |
| JP\_Sales | -0,3087 | 0,1151 | 0,0688 |  |  |  |
| Other\_Sales | 0,0417 | 0,0984 | 0,0690 | 0,0171 |  |  |
| Global\_Sales | -0,7003 | 1,2025 | 0,7133 | 0,2979 | 0,2198 |  |
|  | Year\_of\_Release | NA\_Sales | EU\_Sales | JP\_Sales | Other\_Sales | Global\_Sales |

*Lentelė 4, parodo kovariacijos skaičiavimus, visiems duomenų rinkinio atributams.*

* Koreliacija

Koreliacija išreiškia ryšį, jo stiprumą, tarp dviejų atributų, tačiau šis ryšys yra normalizuotas. Koreliacijos rezultatas yra randamas intervale [-1; 1]. Gavus neigiamą skaičių, galime teigti, jog ryšys yra silpnas, o artėjant prie 1 galim teigti, jog ryšys stiprėja. Todėl naudojant Koreliaciją, galime labai paprastai sudaryti koreliacijos matricą ir ją atvaizduoti naudojant diagramą. Pav. 12.

Su koreliacijos matricos diagrama, galime lengvai pastebėti, jog pardavimai sudaro stiprų ryšį priklausantį vienas nuo kito. Tai pasako, jog jeigu žaidimas yra perkamas viename regione, jis bus perkamas ir kituose regionuose. Labiausiai koreliuoja „Global\_Sales“ su „NA\_Sales“, pasiekiama koreliacijos reikšmę 0.94. Tai jeigu didės „NA\_Sales“ atributo reikšmė, tai didės ir „Global\_Sales“ reikšmė.

Paveikslėlis, kuriame yra ekrano kopija, tekstas, kvadrato formos, Stačiakampis

Automatiškai sugeneruotas aprašymas

pav. 12 Koreliacijos matrica

1. Duomenų normalizacija

Normalizaciją naudojame paversti duomenis į daug suprantamesnę formą. Duomenys yra talpinami į intervalą [0; 1]. Intervale esančius duomenis daug lengviau matyti suprasti ir naudoti. Pav. 13.

Paveikslėlis, kuriame yra tekstas, dokumentas, meniu, popierius

Automatiškai sugeneruotas aprašymas

pav. 13 Normalizuoti duomenys.

1. Išvados

* Po analizės buvo pastebėta, jog buvo atributų, kurie turėjo daug trūkstamų reikšmių. Nesutvarkius duomenų, galėjo kilti daug problemų dėl trūkstamų duomenų su sekančiais skaičiavimais ir vizualizacijose. Norint išspręsti šitą problemą, teko dalį atributų ir duomenų pašalinti, iš skaičiavimų. Pašalintų įrašų skaičius buvo nedidelis tai atsakymas neturėjo pasikeisti sekančiuose skaičiavimuose.
* Modifikuotus duomenis naudojau sudaryti „Scatter plot“ diagramas, kurios vizualiai parodė, ryšį tarp atributų. Šiam duomenų rinkinyje daug ryšių nebuvo, tačiau kai kurie atributai priklausė vienos nuo kito(„Global\_Sales“ ir „NA\_Sales“).