Paveikslėlis, kuriame yra juodas, tamsa

Automatiškai sugeneruotas aprašymas

**Kauno technologijos universitetas**

Informatikos fakultetas

**P176B101 Intelektikos pagrindai**

**Laboratorinis darbas Nr. 1**

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

**KAUNAS, 2024**

Turinys

[Įvadas 4](#_Toc160098600)

[Duomenų rinkinys 5](#_Toc160098601)

[Duomenų rinkinio kokybės analizė 6](#_Toc160098602)

[Tolydinis tipas 6](#_Toc160098603)

[Kategorinis tipas 7](#_Toc160098604)

[Atributų grafikai 8](#_Toc160098605)

[Tolydinio tipo histogramos 8](#_Toc160098606)

[Kategorinio tipo stulpelinės diagramos 12](#_Toc160098607)

[„Scatter plot“ ir SPLOM diagramos 15](#_Toc160098608)

[Kovariacija ir koreliacija 18](#_Toc160098609)

[Kovariacija 18](#_Toc160098610)

[Koreliacija 19](#_Toc160098611)

[Duomenų normalizacija 20](#_Toc160098612)

[Išvados 21](#_Toc160098613)

Paveiksliukai

[pav. 1 „averageUserRating“, „averageUserRatingForCurrenVersion“ atributų analizė 7](#_Toc160098579)

[pav. 2 „fileSizeBytes“, „price“, „userRatingCount“ atributų analizė 8](#_Toc160098580)

[pav. 3 "artistName", "ContentAdvisoryRating", "isGameCenterEnabled" atributų analizė 9](#_Toc160098581)

[pav. 4 "averageUserRating" atributo reikšmių pasiskirstymas 10](#_Toc160098582)

[pav. 5 "averageUserRatingForCurrectVersion" atributo reikšmių pasiskirstymas 10](#_Toc160098583)

[pav. 6 "fileSizeBytes" atributo pasiskirstymas 11](#_Toc160098584)

[pav. 7 "minimumOsVersion" atributų pasiskirstymas 11](#_Toc160098585)

[pav. 8 "price" atributo pasiskirstymas 12](#_Toc160098586)

[pav. 9 "userRatingCount" atributų pasiskirstymas 12](#_Toc160098587)

[pav. 10 "isGameCenterEnabled" atributo reikšmių pasiskirstymas 13](#_Toc160098588)

[pav. 11 "contentAdvisoryRating" atributų reikšmių pasiskirstymas 14](#_Toc160098589)

[pav. 12 "artistName" atributų reikšmių pasiskirstymas 15](#_Toc160098590)

[pav. 13 SPLOM diagrama 16](#_Toc160098591)

[pav. 14 ScatterPlot averageUserRating 17](#_Toc160098592)

[pav. 15 ScatterPlot userRatingCount, fileSizeBytes 17](#_Toc160098593)

[pav. 16 ScatterPlot minimumOSVersion, price 18](#_Toc160098594)

[pav. 17 Atributų kovariacijos HeatMap 19](#_Toc160098595)

[pav. 18 Koreliacijos HeatMap 20](#_Toc160098596)

[pav. 19 Normalizuotų duomenų dalis 21](#_Toc160098597)

# 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.

# Duomenų rinkinys

Laboratorinio darbo atlikimui reikia pasirinkti duomenų rinkinį, kuris turėtų ne mažiau 500 įrašų, bei turėtų ne mažiau kaip 8 stulpelius. Atributus turi sudaryti bent vienas tolydinio bei kategorinio tipo atributas. Duomenų rinkinius galėjome rinktis iš trijų svetainių:

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

Laboratoriniui darbui atlikti pasirinkau duomenų rinkinį: *GameSphere 🕹️: 2000 App Store Insights & Ratings*

<https://www.kaggle.com/datasets/kanchana1990/gamesphere-2000-app-store-insights-and-ratings>

Duomenų rinkinį sudaro 12 stulpelių bei 2000 įrašų, iš kurių 5 atributai yra kategoriniai ir 7 tolydiniai.

Duomenų rinkinio atributai:

* *artistName* – Žaidimų studijos pavadinimas
* *averageUserRating* – Vidutinis žaidėjų įvertinimas
* *averageUserRatingForCurrentVersion* – Vidutinis žaidėjų įvertinimas dabartinei žaidimo versijai
* *contentAdvisoryRating* – Amžiaus cenzo reitingas
* *fileSizeBytes* – žaidimo dydis baitais
* *isGameCenterEnabled* – True/False reikšmės ar suderinamas su apple GameCenter
* *minimumOsVersion* – minimali operacinės sistemos versija
* *price* – žaidimo kaina
* *primaryGenreId* – pagrindinis žaidimo žanro id (nebus naudojamas)
* *releaseDate* – išleidimo data (nebus naudojama)
* *trackName* – žaidimo pavadinimas (nebus naudojama)
* *userRatingCount* – Įvertinimų kiekis

# Duomenų rinkinio kokybės analizė

Duomenys nėra tobuli, todėl pasitaiko įvairių defektų, kurie gali sutrukdyti išgauti korektiškus rezultatus. Tam, kad išvengtume netinkamų rezultatų turime atlikti duomenų rinkinio kokybinę analizę. Šios analizės pagalba rasime defektus, kuriuos spręsime vėlesniuose etapuose.

# 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.

A screenshot of a computer

Description automatically generated

pav. 1 „averageUserRating“, „averageUserRatingForCurrenVersion“ atributų analizė

A screenshot of a computer screen

Description automatically generated

pav. 2 „fileSizeBytes“, „price“, „userRatingCount“ atributų analizė

Tolydinio tipo duomenų rinkinio kokybės analizės rezultatai matomi paveiksliukuose :

pav 1. Ir pav. 2

Kaip ir anksčiau buvo paminėta, šiame duomenų rinkinyje yra 2000 reikšmių, jos buvo patikrintos ir nebuvo rasta trūkstamų reikšmių.

Pagal gautas reikšmes verta paminėti, jog didžioji dalis „*price*“ atributo reikšmių buvo 0, dėl to pirmo ir trečio kvartilio reikšmės yra lygios/ artimos 0.

# 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.

A screenshot of a computer

Description automatically generated

pav. 3 "artistName", "ContentAdvisoryRating", "isGameCenterEnabled" atributų analizė

Kategorinio tipo duomenų rinkinio kokybės analizės rezultatai matomi paveiksliuke pav. 3.

Kaip ir su tolydinio tipo atributais, kategorinio tipo atributų tuščių nebuvo rasta, dėl to įrašų bendras kiekis visgi išlieka 2000.

# Atributų grafikai

# Tolydinio tipo histogramos

Tolydinio tipo atributų reikšmėms atvaizduoti puikiai tinka histogramos. Stulpelių skaičius buvo automatiškai parenkamas Python kalbos įskiepio MatPlotLib. Toliau esančiuose paveiksliukuose parodoma tolydinio tipo atributų reikšmių pasisikirstymai.

A graph of a user rating

Description automatically generated

pav. 4 "averageUserRating" atributo reikšmių pasiskirstymas

Kaip matome pav. 4, didžioji dauguma žaidėjų žaidimus vertino pozityviai. Histograma pasislinkus į dešinę.

A graph of a person with red bars

Description automatically generated with medium confidence

pav. 5 "averageUserRatingForCurrectVersion" atributo reikšmių pasiskirstymas

Kaip matoma pav. 5, žaidėjų vidutiniai vertinimai nelabai skiriasi nuo bendro vidutinio įvertinimo.

A graph of a number of files

Description automatically generated

pav. 6 "fileSizeBytes" atributo pasiskirstymas

Kaip matoma pav. 6, Didžioji dauguma žaidimų užima iki 1 milijardo bitų.

A graph with red bars

Description automatically generated

pav. 7 "minimumOsVersion" atributų pasiskirstymas

Kaip matoma pav. 7, minimali OS versija yra versija 12.

A graph with red and black lines

Description automatically generated

pav. 8 "price" atributo pasiskirstymas

Kaip matoma pav. 8, praktiškai visi žaidimai sąraše yra nemokami.

A red graph with black text

Description automatically generated

pav. 9 "userRatingCount" atributų pasiskirstymas

Kaip matoma pav. 9, vidutinis žaidimas turėjo virš 10000 įvertinimų.

# Kategorinio tipo stulpelinės diagramos

Kategorinio tipo reikšmių atvaizdavimui bus naudojamos stulpelinės diagramos.

A green rectangular bar graph

Description automatically generated

pav. 10 "isGameCenterEnabled" atributo reikšmių pasiskirstymas

Kaip matome pav. 10, apie 70 procentų žaidimų neturi *Game Center* parengimo.

A bar graph with green rectangular bars

Description automatically generated

pav. 11 "contentAdvisoryRating" atributų reikšmių pasiskirstymas

Kaip matoma pav. 11, didžioji dalis žaidimų tinka jaunesnio amžiaus asmenims 4-12 metų.

A graph with text overlay

Description automatically generated

pav. 12 "artistName" atributų reikšmių pasiskirstymas

Kaip matoma pav. 12, yra kelios įmonės, kurios yra išleidę daugiau nei 40 žaidimų šiame sąraše.

# „Scatter plot“ ir SPLOM diagramos

"*Scatter plot*" bei *SPLOM* yra duomenų vizualizacijos metodai, kurie naudojami įvertinti ryšį tarp dviejų ar daugiau atributų.

Galime pastebėti, kad didžioji dauguma atributų tarpusavyje ryšių neturi, tačiau „*averageUserRating*“ bei „*averageUserRatingForCurrectVersion*“ yra beveik identiški dėl to, kad iš esmės vidutiniai įvertinimai skaičiuojami pridėjus visų versijų įvertinimus, ne tik naujausios versijos. Tai parodo pav. 13 bei pav. 14.

A screenshot of a graph

Description automatically generated

pav. 13 SPLOM diagrama

A graph with a red line

Description automatically generated

pav. 14 ScatterPlot averageUserRating

A graph with a red line and a line

Description automatically generated

pav. 15 ScatterPlot userRatingCount, fileSizeBytes

A graph with a red line

Description automatically generated

pav. 16 ScatterPlot minimumOSVersion, price

# Kovariacija ir koreliacija

# Kovariacija

Kovariacija parodo dviejų atributų ryšio stiprumą vienas tarp kito.

Kovariaciją tarp atributų atvaizduoju kovariacijos *HeatMap* (pav. 17)

A screenshot of a computer

Description automatically generated

pav. 17 Atributų kovariacijos HeatMap

# Koreliacija

Koreliacija išreiškia ryšį, jo stiprumą, tarp dviejų atributų, tačiau šis ryšys yra normalizuotas. Koreliacijos rezultatas yra randamas intervale [-1; 1]. Jeigu koreliacijos rezultatas yra mažesnis nei 0, vadinasi – ryšys silpnas, jeigu virš 0 – stiprus. Koreliacijai atvaizduoti taip pat pasitelkiu *HeatMap*. (pav. 18)

A screenshot of a graph

Description automatically generated

pav. 18 Koreliacijos HeatMap

# Duomenų normalizacija

Duomenų normalizaciją į intervalą [0;1] atliekame tada, kai norime turimus duomenis paversti į suprantamesnę formą. Į šį intervalą sutalpinus duomenis, juos yra paprasčiau suprasti bei naudoti. Pav. 19

A table of numbers and letters

Description automatically generated

pav. 19 Normalizuotų duomenų dalis

# Išvados

* Analizės metu buvo pastebėta, kad nebuvo trūkstamų reikšmių atributuose, tačiau buvo kategorinių reikšmių, kurios nebuvo suprantamos Python bei Excel programinėms įrangoms (Japonų, Kinų raidės). Kadangi tai daug įtakos nesudarė, problemų nekilo, tačiau jeigu tokių reikšmių būtų sutikta tolydinio tipo atributų reikšmėse, būtų kilę problemų su sprendimu.
* Sunormalizavus visas atributų reikšmes outlier skaičius nepakito, tačiau sunormalizavus ir sudėjus į vienodą skalę outlier išsišokimai neatrodo toki dideli.

Gynimo klausimai:

Ar po z score normalizacijos bus daugiau outlierių?

Mano duomenų rinkinyje jų skaičius nesumažėjo

A screenshot of a phone

Description automatically generated