VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
EKONOMETRIJOS BAKALAURO STUDIJŲ PROGRAMA

Roberta Morkūnaitė

Audiovizualinių kūrinių autorių teisių pažeidimų internete statistinė analizė

Statistical Analysis of Online Audiovisual Production Copyright Infringement

Bakalauro baigiamasis darbas

Darbo vadovas Doc., Dr., Gediminas Murauskas

VILNIUS
2020

Turinys

Audiovizualinių kūrinių autorių teisių pažeidimų internete statistinė analizė

Santrauka

Vykstant greitai technologijų raidai viena pažeidžiamiausių pramoginių sričių tampa kino industrija. Įrašų studijų išleidžiami audiovizualiniai kūriniai dažnai nelegaliai platinami internete, kuriame vartotojai gali lengvai, greitai ir nemokamai juos parsisiųsti. Taip prarandama ne tik didelės pajamos, bet ir pažeidžiamos autorių teisės. Todėl svarbu identifikuoti grėsmingiausius pasaulio regionus, susidaryti potencialaus pažeidėjo "portretą", nuspėti pikus, kad būtų kuo efektyviau mažinamas pažeidimų skaičius ir užtikrinama autorių teisių apsauga. Šiame darbe analizuojami dviejų filmų P2P technologijos pažeidimų skaičiaus šalyse duomenys ir taikomas laiko eilučių klasterinės analizės metodas bei sudarinėjami logistinės regresijos modeliai. Grupuojamos šalys pagal pažeidimų skaičiaus augimo/mažėjimo tendencijas, nagrinėjama modelių suteikta informacija apie pagrindines priklausomybes, sąryšius tarp pažeidimų ir ekonominių bei socialinių rodiklių.

Raktiniai žodžiai : Audiovizualiniai kūriniai; Autorių teisės; P2P technologija; Logistinės regresijos modelis; Klasterinė laiko eilučių analizė;

Statistical Analysis of Online Audiovisual Production Copyright Infringement

Summary

With the rapid development of technology, the film industry is becoming one of the most vulnerable areas of entertainment. Audiovisual works published by recording studios are often illegally leaked online, where users can download them easily, quickly and free of charge. This not only results in a significant loss of revenue, but also in copyright infringement. It is therefore important to identify the most threatening regions of the world, to create a "portrait" of a potential infringer, to anticipate peaks in order to reduce the number of infringements as effectively as possible and to ensure copyright protection.In this work, the data of P2P technology violations of two films in the countries are analyzed using time series cluster analysis method and logistic regression models. Groups of countries according to the increasing/decreasing trends in the number of violations and information provided by the models on the main dependencies between violations and economic and social indicators are analyzed.

Key words : Audiovisual works; Copyrights; P2P technology; Logistic regression model; Time series clustering analysis,

1 Įvadas

Šiuolaikinė visuomenė neretai, kaip laisvalaikio praleidimo būdą, renkasi mėgstamų filmų, serialų peržiūrą. Tobulėjant technologijoms, tampa vis lengviau pasiekti norimą turinį be didelių pastangų, neišeinant iš namų, todėl dažnai autoriai, ar oficialūs turinio transliuotojai susiduria su nelegaliu turinio platinimu internete. Platinant, ar naudojantis nelegaliu turiniu yra pažeidžiamos turtinės ir moralinės autorių teisės.

Viena populiariausių technologijų, siekiant parsisiųsti didesnio formato nelegalius failus, yra torentai, kurie pagrįsti vadinamu P2P (angl.: Peer-to-peer) tinklo modeliu, kuomet vartotojai, privatūs asmenys, dalinasi rinkmenomis tarpusavyje.

Siekiant mažinti nelegalaus turinio internete mastą ir pasiruošti galimiems pažeidimų pikams ateityje, svarbu ištirti suaktyvėjusio piratavimo priežastis bei faktorius. Atrasti ekonominių rodiklių sąryšius su fiksuojamu pažeidimų skaičiumi bei demografinę priklausomybę, kad būtų lengviau identifikuoti galimą pažeidėją ir prognozuoti apimtis ateityje.

Darbo tikslas – ištirti priklausomybes, sąryšius tarp pažeidimų ir ekonominių bei demografinių rodiklių, audiovizualinių kūrinių piratavimo internete apimtims prognozuoti; Identifikuoti potencialius pažeidėjus;

Norint pasiekti šį tikslą būtina išspręsti **uždavinius**:

1. Išanalizuoti piratavimo internete torentų technologijos duomenis, pastebėti nelegalaus turinio sklaidos tendencijas pasaulyje;
2. Panagrinėti demografinius bei ekonominius rodiklių duomenis, naduojantis matematiniais metodais įvertinti jų sąryšį su pažeidimų skaičiumi šalyse;
3. Apibendrinti tyrimo rezultatus ir parengti išvadas.

Atliekant šį tyrimą, naudoti **metodai** ir **modeliai**:

- Statistinė duomenų analizė (Markmonitor įmonės duomenų bazė, Euromonitor duomenų bazė);
- Mokslinės literatūros analizė;
- Dispersinė analizė ANOVA;
- Klasterinė laiko eilučių analizė;
- Logistinės regresijos modelis.

2 Piratavimas, temos aktualumas šių dienų visuomenėje

Tobulėjant technologijoms ir keičiantis būdams įsigyti filmus, ar serialus internete, didėja autorių teisių pažeidimų mastas. Visuomenė nuo fizinių laikmenų, tokių kaip DVD, ar CD kompaktinių diskų, perėjo prie audiovizualinio turinio įsigyjimo virtualioje erdvėje. Vis dažniau kompanijos, kuriančios filmus, serialus domisi ir ieško efektyvaus būdo, kuo greičiau nustatyti pažeidimus ir juos pašalinti.

Kova su piratavimu yra nauja sritis, kurios vystymosi greitis didelis, todėl svarbu prisitaikyti prie vartotojų elgsenos ir kuo greičiau identifikuoti galimus pažeidimus.

Nagrinėjant jau atliktas analizes, galima atsižvelgti į ankstesnes, dažniausiai susiduriamas problemas, susijusias su naujomis turinio platinimo platformomis, būdais siekiant pašalinti nelegalų turinį ir taip užkirsti kelią piratavimui internete.

2018 metų Amsterdamo universiteto profesorių analizės [1] tikslas, įvertinti piratavimo poveikį legalaus turinio transliuotojams mastą 13-oje pasaulio šalių bei identifikuoti vartotojus, linkusius naudotis nelegaliu turiniu. Straipsnyje įžvelgiami teigiami piratavimo poveikiai legaliam turiniui - pristatant tam tikrus kūrinius, filmų žanrus, sukuriama papildoma paklausa ateičiai, taip pat ir neigiami - piratavimas skatina neieškoti, kur įsigyti legalų turinį, nesilankyti kino teatruose ir taupant laiką, filmus/serialus/tiesiogines transliacijas, žiūrėti lengvai prienamuose tinklalapiuose.

Straipsnyje naudojami duomenys gauti apklausus 35 tūkst. respondentų, tarp kurių 7 tūkst. nepilnamečių iš 13 šalių (Europos, Jungtinių Amerikos Valstijų ir Azijos), taip pat buvo naudojami šalių duomenys gauti iš Pasaulio Banko.

Naudojantis Pasaulio Banko duomenimis apie šalių populiaciją, pasiskirstymą amžiaus grupėse ir pajamas vienam gyventojui, nustatyta, kad piratavimas yra paplitęs jaunesnėje amžiaus grupėje, dažniausiai tarp vyrų. Įžvelgiama koreliacija tarp pirkimo galios ir piratavimo internete.

Per pastaruosius 3 metus išaugo prenumeratų pirkimas tokiose platformose, kaip Netflix, ar turinio įsigyjimas naudojantis Apple TV platformomis. Įžvelgiama mažėjantis fizinis vaizdo-garso turinio pardavimas visose šalyse, tačiau skaitmenins pardavimų augimas lėmė bendrą gryną audiovizualinių kūrinių pardavimų augimą 2014-2017 metais, kas įrodo, kad vis daugiau vartotojų persikelia į virtualią erdvę. Apklausos metu, dažniausia priežastis, siunčiantis nelegalų turinį internetu, buvo nurodyta kaina. Tam tikrose šalyse, pavyzdžiui Indijoje, ar Japonijoje nurodoma vaizdo ir garso kokybės priežastis, taip pat lengvas ir greitas pasiekiamumas Honkonge. Be to, pastebima, kad nelegaliai pasiekiamų audiovizualinių kūrinių vartojimas dvigubai didesnis, nei legalių.

Pasak MUSO straipsnyje minimo tyrimo rezultatų [2] 2017 metais nustatyta daugiau kaip 300 bilijonų apsilnakymų piratiniuose puslapiuose ir pastebimas 1.6 procento augimas lyginant su 2016 metais. Daugiausiai lankytojų piratiniuose puslapiuose yra iš Rusijos,

Indijos ir Brazilijos. Populiariausia nelegalaus turinio žiūrėjimo platforma yra tinklalapiai, kuriuose filmai tranliuojami tiesiogiai, be poreikio parsisiųsti.

Tai įrodo, jog piratavimuas internete šiuo laikotarpiu yra suaktyvėjas ir kelia vis didesnį susidomėjimą audiovizualinių kūrinių autoriams. Tikima, kad tobulėjant technologijoms, gali atsirasti vis daugiau būdų lengvai pasiekti nelegalų turinį internete.

2.1 Piratavimo technologija

Analizėje naudojami P2P technologijos duomenys. P2P technologija, tai failų apsikeitimas tarp vartotojų (iš kompiuterio į kompiuterį). Daugelis tikisi, kad naudojimasis P2P technologija piratavimo tikslais ateityje sumažės dėl vis didėjačių nelegalaus turinio transliuotojų tinklalapiuose, tačiau ši platforma vis dar nepraranda populiarumo, todėl verta dėmesio ir išsamesnės analizės [3].

2017 m. užfiksuota, kad pasiskirstymas tarp šių dviejų platformų - P2P ir video transliuotojų internetu yra toks: P2P tinklalapiai globaliai sudaro 70proc., o video transliuotojai internetu 30proc. Stebint pasaulines tendencijas, svarbu teisingai įvertinti pažeidimų mastą įvairiose platformose ir neleisti plisti nelegaliam turiniui internete, bei per anksti nenuvertinti senos technologijos platformų [3].

3 Taikomi metodai ir modeliai

Skyriuje trumpai aptariami audiovizualinių kūrinių teisių pažeidimų internete analizei taikyti metodai ir modeliai:

- dviejų rūšių (parametrinė ir neparametrinė) dispersinės analizės - siekiama išsiaiškinti sąryšį tarp pažeidimų skaičiaus 100 tūkst. gyventojų šalyse ir pajamų grupės į kurią jos patenka;
- laiko eilučių klasterinė analizė - naudojantis dieniniais šalių pažeidimų duomenimis, nustatyti šalių pažeidimų tendencijas ir tarpusavio sąsajas, suskirstant tiriamus objektus (šalis) į klasterius, pagal požymių panašumą;
- logistinės regresijos modelis - pagrindinių piratavimo masto veiksnių identifikavimas ir jų statistinio reikšmingumo vertinimas, naudojantis binarine (angl. binary) logistine regresija.

3.1 Dispersinė analizė

3.1.1 Dispersinė analizė ANOVA

Analizei taikomas **Welch F kriterijus** (angl.: Welch ANOVA). Welch ANOVA naudojama, kuomet turime normalius duomenis, bet skirtingas dispersijas. Duomenys turi tenkinti prielaidas:

1. Grupės nepriklausomos tarpusavyje;
2. Grupės tenkina normalumo sąlygą;
3. Grupių dipersijos skirtingos.

Imtis suskaidoma į keletą grupių pagal vieną požymį ir tikrinama, ar skirtingų grupių vidurkiai statistiškai reikšmingai skiriasi. Kitaip sakant, yra tikrinama hipotezė:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_k \\ H_1 : \mu_i \neq \mu_s \end{cases}$$

μ - grupių vidurkiai, k - lyginamų grupių skaičius, μ_i, μ_s - bet kurios grupės vidurkis iš visų k grupių.

Sprendimo priėmimo taisyklė pasirinktam reikšmingumo lygmeniui α:

$$F_{welch} \leqslant F_{1-\alpha;k-1,1/\Lambda)},\ priimame\ H_0;$$

$$F_{welch} > F_{1-\alpha;k-1,1/\Lambda)},\ priimame\ H_1.$$

Naudojama Welch F-testo statistika F_{welch}. Esminis skirtumas nuo F-statistikos naudojamos atliekant vienfaktorinę ANOVA, kad kiekvienai grupei priskiriami svoriai (w_i) sumažinti heteroskedastiškumo poveikį analizės rezultatams. F_{welch} statistika apskaičiuojama:

$$F_{welch} = \frac{SSTR_{welch}/(k-1)}{1 + \frac{2\Lambda(k-2)}{3}} \sim F_{k-1,1/\Lambda},$$

parametras Λ, toliau naudojamas laisvės laipsniams skaičiuoti:

$$\Lambda = \frac{3\sum_{i=1}^{k} \frac{\left(1-\frac{w_i}{\sum_{i=1}^{k} w_i}\right)^2}{n_i - 1}}{k^2 - 1}.$$

$SSTR_{welch}$ kvadratų suma (angl. treatment sum of squares) aiškinanti, kiek skiriasi tiriamų grupių vidurkiai:

$$SSTR_{welch} = \sum_{i=1}^{k} w_i(\overline{Y_i} - \overline{Y}_{welch})^2.$$

$\overline{Y}_{welch}$ bendras visų grupių imties vidurkis skaičiuojamas:

$$\overline{Y}_{welch} = \frac{\sum_{i=1}^{k} w_i \overline{Y_i}}{\sum_{i=1}^{k} w_i},$$

w_i - i-tosios grupės svoris, nustatomas pagal grupės imties dydį n_i ir stebėjimų variacijos s_i^2,

$$w_i = \frac{n_i}{s_i}.$$

[8] Kaip ir minėta, pagrindinis tikslas renkantis Welch F-testą yra heteroskedastiškumo problemos poveikio mažinimas, priskiriant grupėms svorius.

Welch ANOVA naudojama nustatyti, ar bent dviejų grupių (iš bednro grupių skaičiaus k) vidurkiai skiriasi, bet jas identifikuoti taikomi **aposterioriniai (Post-Hoc) kriterijai**. Dažniausiai - Tukey HSD arba Welch ANOVA atveju - **Games Howell kriterijus**, kuris parodo, kurių grupių vidurkiai statistiškai reikšmingai skiriasi.

Analizėje naudojami abu testai ir lyginami jų rezultatai siekiant tesingai formuoti išvadas.

Games Howell kriterijus yra Tukey-Kramer krierijaus atitikmuo, kai duomenys nėra homogeniški. Tai yra t-testas su Welch laisvės laipsniais.

Tikrinamos hipotezės apie dviejų grupių vidurkių lygybę:

$$\begin{cases} H_0 : \mu_i - \mu_j = 0 \\ H_1 : \mu_i - \mu_j \neq 0 \end{cases}$$

μ_i, μ_s - skirtingų grupių vidurkiai.

Sakome, kad grupių vidurkiai statistiškai reikšmingai skiriasi, kai vidurkių skirtumas didesnis už minimalų reikšmingą skirtumą $q_{\sigma,k,df}$[9]:

$$\overline{x_i} - \overline{x_j} > q_{\sigma,k,df},$$

čia standartinės paklaidos σ skaičiuojamos:

$$\sigma = \sqrt{\frac{1}{2}\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{nj}\right)},$$

laisvės laipsniai:

$$df = \frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{nj}\right)^2}{\frac{\left(\frac{s_i}{n_i}\right)^2}{n_i - 1} + \frac{\left(\frac{s_j}{n_j}\right)^2}{n_j - 1}},$$

k - grupių skaičius;

$\overline{x}_i$, $\overline{x}_j$- grupės i ir j vidurkiai, $i, j \in (1, 2, .., k)$;

n_i, n_j - i-tosios ir j-tosios grupių imties dydis, $i, j \in (1, 2, .., k)$;

s_i^2, s_j^2 -i-tosios ir j-tosios grupių dispersijos, $i, j \in (1, 2, .., k)$.

Reikia pabrėžti, kad viskas skaičiuojama kiekvienai grupei atskirai, dėl nevienodų grupių imties dydžių.

3.1.2 Kruskal – Wallis testas

Kruskal - Wallis yra neparametrinis kriterijus analogiškas vienfaktorinei nepriklausomų imčių ANOVA metodui, tikrinantis ne vidurkius, o grupių pasiskirstymą. Taip pat, nenaudojami aposterioriniai (Post-Hoc) kriterijai ir reikalaujama mažiau prielaidų - duomenys neturi būti normaliai pasiskirstę.

Nustatomas tik bendras grupių poveikis priklausomam kintamajam Y. Tikrinama statistinė hipotezė:

$$\begin{cases} H_0 : F_1(x) = F_2(x) = ... = F_k(x) \\ H_1 : \exists \textit{ bent viena nelygybė} \end{cases}$$

$F_k(x)$ - k-tosios grupės pasiskirstymas, $x = x_{k1}, x_{k2}, ..., x_{kn_k}$, kur k - grupių skaičius. Tegul $x_{i,j}$ - i-tosios grupės j-tasis stebėjimas, $i = 1, 2, ..., k$ ir $j = 1, 2, ..., n_k$.

Naudojama sprendimo priėmimo taisyklė:

$$H > \chi^2_{1-\alpha,k-1} - \textit{atmetame } H_0,$$

čia α pasirinktas reikšmingumo lygmuo, o $\chi^2_{1-\alpha,k-1}$ chi-kvadrato skirstinio su k-1 laisvės laipsniais $(1 - \alpha)$ kvantilis.

H statistika atrodo taip:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1),$$

čia $N = \sum_{i=1}^{k} n_i$ - nuo mažiausio iki didžiausio išrikiuotų stebėjimų rangas, didžiausio stebėjimo rangas - N;

$R_i = \sum_{j=1}^{n_i} R_{ij}$, kur R_{ij} stebėjimo x_{ij} rangas, o R_i yra i-tosios grupės rangų suma [10].

Įsitikinus, kad Kruskal–Wallis testas statistiškai reikšmingas, galima daryti išvadas apie grupių pasiskirstymus, tai nėra labai informatyvus testas, todėl atsiranda poreikis ieškoti galimybių atlikti aposteriorinius testus. Tai galima padaryti naudojantis R programos funkcija *dunnTest()* [6], kuri atlieka Dunn testą, nusakantį skirtingų imčių grupių skirtumų statistinį reikšmingumą.

Dunn testas - neparametrinis kriterijus, atliekantis porinius vidurkių palyginimus skirtingo dydžio grupių imtims. Naudojama z-statistika. H_0: tikimybė, kad atsitiktinis stebėjimas pirmoje grupėje yra didesnis už atsitiktinį stebėjimą antroje grupėje lygi 0.5 [11]. Dunn z-testo statistika, lyginant grupes, naudoja Kruskal-Wallis testo metu gautus vidutinius kiekvienos grupės rangus.

3.2 Laiko eilučių klasterinė analizė

Laiko eilučių klasterinė analizė pradedma pasirenkant:

1. Atstumo mato tarp panašių objektų skaičiavimo metodą;
2. Klasterizavimo strategiją;
3. Klasterių skaičių.

3.2.1 Atstumo matas

Grupuojant stebėjimus, reikalingas matas apskaičiuoti atstumą, ar panašumą tarp kiekvienos stebėjimo poros. Kitaip tariant, reikia rasti atstumo skirtumų matricą. Pagrindiniai atstumo matai:

- Euklido atstumas (angl. Euclidean distance) - tiesiausias atstumas tarp dviejų taškų. Nėra tinkamas, kuomet duomenys turi didelius pikus ir laiko eilučių ilgis nesutampa.

$$dist_e(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

- Kanberos atstumas (angl.Canberra distance) - atstumas, skaičiuojamas pagal laiko eilučių tendecijų panašumus, bet ne pagal artumą. Jautrus reikšmėms arti 0, jei skaitiklyje, ar vardiklyje gaunamas 0 atstumas tarp stebėjimų išbraukiamas iš sumos ir traktuojamos, kaip NA reikšmė.

$$dist_c(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_j| + |y_j|},$$

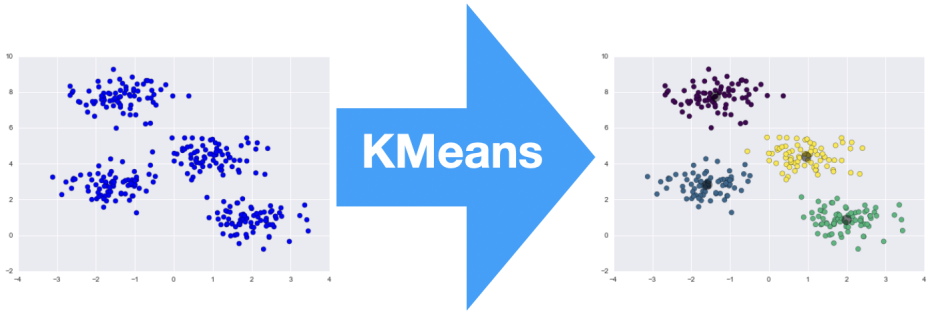
čia $dist_e(x, y)$ - euklidinis atstumas tarp dviejų laiko eilučių $x = (x_1, x_2, ..., x_k)$ ir $y = (y_1, y_2, ..., y_l)$, kur $x \in \mathbb{R}^k$, $y \in \mathbb{R}^l$, $k, l \in \mathbb{Z}$, n - bendras stebėjimų skaičius laiko eilutėje, $dist_c(x, y)$ - Kanaberos atstumas tarp dviejų laiko eilučių taškų.

3.2.2 Klasterizavimo strategija

Skiriamos pagrindinės dvi klasterizavimo metodikos:

- **Dalijimo** arba nehierarchinis metodas (angl. partitioning clustering/non-hierarchical clustering) (žr. pav. 1)[15]. Šio tipo populiariausias K-vidurkių klasterizavimo metodas (angl. k-means clutering). Metodo privalumai - greitis ir paprastumas, todėl gali būti taikomas dideliems duomenų rinkiniams. To priežastis - iš anksto nurodomas klasterių skaičius. Algoritmo trūkumai - jautrus išskirtims ir kiekvieną kartą gali duoti skirtingus rezultatus. Dažnai, kaip trūkumas, ypač nagrinėjant laiko eilutes, nurodomas išanksinis klasterių skaičiaus pasirinkimas. Dalijimo metodas turėtų būti naudojamas mažų dimesijų ir aiškiai atskirtiems duomenims, klasterių skaičiaus nustatymui lengvinti.

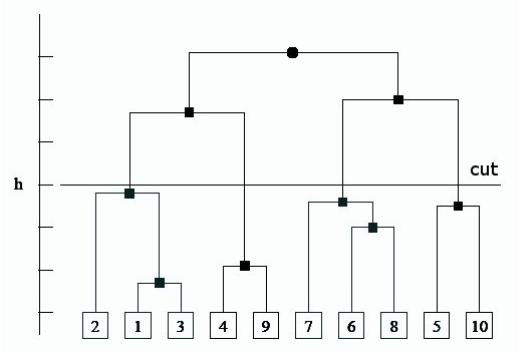
1 pav.: K-vidurkių metodas



- **Hierarchinis** klasterizavimas (angl. hierarchical clustering).

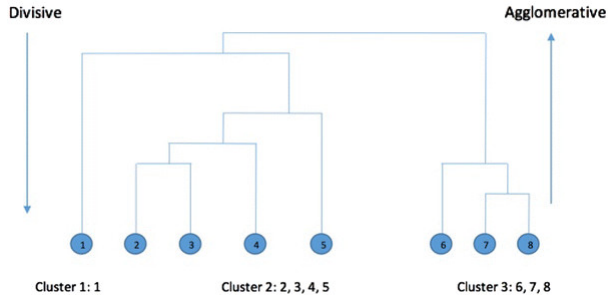
  Skirtinagai nuo dalijimo metodo, kur nesikertantys duomenys yra suskirtomi į grupes, čia visi duomenys priskiriami vienam klasteriui, kuris skaidomas į kitus klasterius ir yra vaizduojami dendogramomis. Tuomet, klasterių skaičius nustatomas atliekant pjūvius dendogramoje (žr. pav. 2)[16]. Paveikslėlyje matoma, kaip pjūviu b yra sukonstruojami 4 klasteriai. Šis metodas dažnai taikomas laiko eilučių analizei, nes nereikia iš anksto nustayti klasterių skaičiaus ir lengviau pamatyti tendencijas dendogramoje. Toliau išsamiau aptariama šio metodo veikimas.

2 pav.: Hierarchinis klasterizavimas



Hierarchinis klasterizavimo metodas skirstomas į jungimo (angl. agglomerative method) ir skaidymo (angl. divisive method) klasterizavimo tipus (žr. pav. 3)[17].

3 pav.: Hierarchinio klasterizavimo tipai



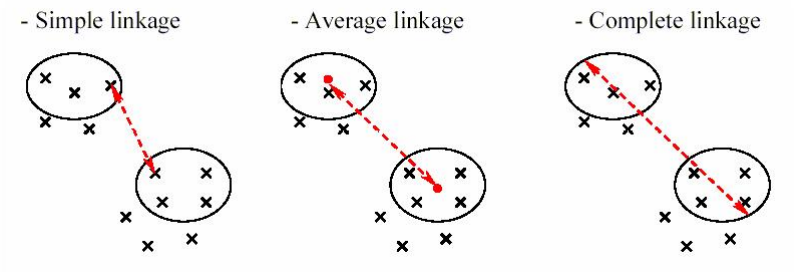
Jungimo metodas visus stebėjimus laiko atskirais klasteriais. Vis prijungiant artimiausius stebėjimus, formuojami didesni klasteriai, kol išskiriamas vienas bendras stebėjimų klasteris. Jugti klasterius galima trim būdais (žr. pav. 4)[18]:

- Artimiausio kaimyno metodu (angl. single linkage) - minimizuojant atstumą tarp stebėjimų porų.
$$d(X,Y) = min(d(x_u, y_j), d(x_i, y_j))$$
- Tolimiausio kaimymo metodu (angl. complete linkage) - maksimizuojant atstumą tarp stebėjimų porų.
$$d(X,Y) = max(d(x_u, y_j), d(x_i, y_j))$$
- Klasterio centrų jungimo metodu (angl. average linkage) - randat panašiausius vidurkius.
$$d(X,Y) = d(\overline{X}, \overline{Y}) = \sum_{x_i \in X} \sum_{y_i \in Y} \frac{d(x_i, y_j)}{n_i, n_j}$$
- Mažiausios dispersijos metodas (angl. Ward method) - alternatyva artimiausio kaimyno metodui, tinkanti kiekybiniams kintamiesiems, matuojanti kiek padidės kvadratų

suma sujungus du klasterius.

$$d(X,Y) = \frac{x_i y_j}{x_i + y_j} \|X' - Y'\|^2$$

4 pav.: Hierarchinio klasterių jungimo būdai



Čia X, Y - skirtingi klasteriai, $x = (x_1, x_2, ..., x_k)$- klasterio X stebėjimai, $y = (y_1, y_2, ..., y_l)$ - klasterio Y stebėjimai, kur $x \in \mathbb{R}^k$, $y \in \mathbb{R}^l$, $i, j, u, v \in \mathbb{Z}$, $d(X.Y)$ - atstumas tarp klasterių, $\overline{X}, \overline{Y}$ - klasterių požymio vidurkiai, X', Y'- klasterių X ir Y centrai, stebėjimų vidurkiai apskaičiuojami kiekvienam klasteriui.

Skaidymo metodas yra atvirkščias jungimo metodui. Visi stebėjimai laikomi vienu klasteriu ir pakartotinai atskiriant tolimiausius stebėjimus, suskaidomi į mažesnius klasterius.

3.2.3 Klasterių skaičiaus pasirinkimas

Pagrindinis skirtumas tarp nehierarchinių ir hierarchinių metodų - klasterių skaičiaus pasirinkimas iš anksto. Todėl svarbu turėti įrankius, kurie padėtų nustatyi optimalų klasterių skaičių.

Dažniausiai naudojami klasterių skaičiaus parinkimo matai:

- **Elbow** metodas - pagrindinė idėja minimizuoti variacijos reikšmę tarp dviejų klasterių, skaičiuojant bendrą klasterių kvadratų sumą (angl.total within-cluster sum of square).

  $$min\left(\sum_{k=1}^{k} W(C_k)\right),$$

  čia C_k - k-tasis klasteris, $W(C_k)$ - bendra klasterių kvadratų suma, $k \in Z$ - klasterių skaičius.

- **Silhouette** metodas - matuoja, kaip gerai kiekvienas stebėjimas tinka priskirtam klasteriui. Nusako klasterio kokybę. Optimalus klasterių skaičius yra tas, kuris maksimizuoja vidutinę Shilhouette reikšmę, apskaičiuojamą:

  $$Coef_{silh} = \frac{C_i - C_n}{max(C_i, C_n)},$$

čia C_i - i-tojo klasterio vidutinis atstumas tarp pasirinkto stebėjimo ir kitų to paties klasterio stebėjimų, C_n - vidutinis n-tojo artimiausio klasterio atsumas, i ir n - klasterių skaičius. Šis koefcientas yra intervale [-1,1]. Artima reikšmė 1-ui rodo, kad stebėjimui parinktas geras (artimiausias) klasteris, jei reikšmė -1 - tuomet stebėjimui priskirtas neteisingas klasteris.

3.3 Logistinės regresijos modelis

Analizėje naudojama binarinė logistinė regresija. Regresijos priklausomas **dvireikšmis** kintamasis Y (pažeidimų skaičius internete) įgyja reikšmes 0 arba 1. Modelis užrašomas taip:

$$P(Y = 1) = \frac{e^{z(x)}}{1 + e^{z(x)}},$$

$z(x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m$, kur $x_1, x_2, ..., x_m$ - nepriklausomi kintamieji, m - nepriklausomų kintamųjų skaičius.

Sudarinėjant logistinę regresiją, svarbu mokėti interpretuoti gautus rezultatus, todėl reikia žinoti keletą sąvokų:

Galimybė (odds) įvykti įvykiui, kai priklausomas kintamasis Y įgis reikšmę 1 apskaičiuojama tikimybių santykiu:

$$\frac{P(Y = 1)}{1 - P(Y = 1)}.$$

Galimybių santykis (angl. odds ratio), kuris apskaičiuojamas:

$$OR = e^{\beta(k)},$$

čia β_k koeficientas, OR - galimybių santykis.

Galiausiai, naudojantis apskaičiuotu galimybių santykiu, galima vertinti nepriklausomo kintamojo x_k poveikį tikimybių santykiui (galimybei), padidinus jį vienetu [4].

4 Duomenys

4.1 Tyrimo ojektas - pažeidimai

Tyrimo analizei atlikti pasirinkti du filmai: Ready (2019) ir Beast (2019). Visa antrojo filmo Beast informacija, grafikai ir lentelės pateiktos priede 8.1. Siakiant užtikrinti adekvatų palyginimą, filmus stengiamasi parinkti kuo panašesnia charakteristika:

1. Filmų žanras;
2. Pirminė (šalies viduje) išleidimo data;
3. IMDB įvertinimas;
4. Biudžetas.

Verta paminėti, kad filmų kino studijos skiriasi. Pasirinkti filmai tik dalinai atspindi fiksuojamų pažeidimų mastą pasaulyje, nes remiantis dviejų filmų pažeidimų duomenimis, įvertinti tikrąją situaciją ir daryti išvadas apie visą populiaciją, būtų sudėtinga. Be to, kaupti viso pasaulio pažeidimų duomenis duomenų bazėse - neįmanoma. Tačiau šių dviejų filmų žanrų skirtumas ir didelis populiarumas užtikrina, kad analizė būtų kuo informatyvesnė ir atspindėtų įvairaus amžiaus, lyties, ar kultūros auditoriją bei jos elgseną.

Filmų specifikacija aprašyta 1 lenetlėje.

1 lentelė: Filmų specifikacija

Filmas	**Žanras**	**Išleidimo data šalies viduje**	**Bendras IMDB įvertinimas**	**Balsavusių žmonių skaičiaus**	**Pajamos iš kino teatro pasaulyje USD**	**DVD pardavimai šalies viduje USD**	**Blue-ray pardavimai šalies viduje USD**	**Biudžetas USD**
Ready	Komedija Drama	19/07/26	7.7	459 tūkst.	372 mln.	6 mln.	13.6 mln.	90 mln.
Beast	Komedija Veiksmo	19/07/02	7.5	287 tūkst.	1.1 bil.	13.5 mln.	39.6 mln.	160 mln.
Filmų įvertinimo ir kainų santykis	-	-	-	1.59	3.04	2.22	2.90	1.78

IMDB puslapio duomenimis filmo "Ready" įvertinimas yra 0,2 balo didesnis už filmo "Beast" įvertinimą. "Ready" filmo balsuotojų buvo 1,6 karto daugiau, nei filmo "Beast" balsuotojų. Nors pajamos, gautos iš vietinių bei pasaulio šalių kino teatrų ir parduotų DVD/Blue-ray diskų yra 2/3 kartus didesnės nei filmo "Ready" gautos pajamos. Filmo "Beast" biudžetas taip pat 1,8 karto didesnis, nei filmo "Ready".

Pajamų skirtumai gali atsirasti dėl skirtingų DVD/Blue-ray diskų, ar kino teatro bilietų kainų. Dažnai kino teatrai turi fiksuotas kainas pagal amžiaus grupes filmams žiūrėti, todėl spręsti apie surinktas pajamas naudojantis kino teatro kainomis pasaulyje, būtų sudėtinga. Pasidomėjus DVD ir Blue-ray diskų kainomis "Amazon" svetainėje, rastos tokios kainos USD valiuta (žr. lent. 2).

2 lentelė: kainos

Filmas	**DVD kaina**	**Blue-ray kaina**
Ready	16.12	20.73
Beast	8.66	12.39

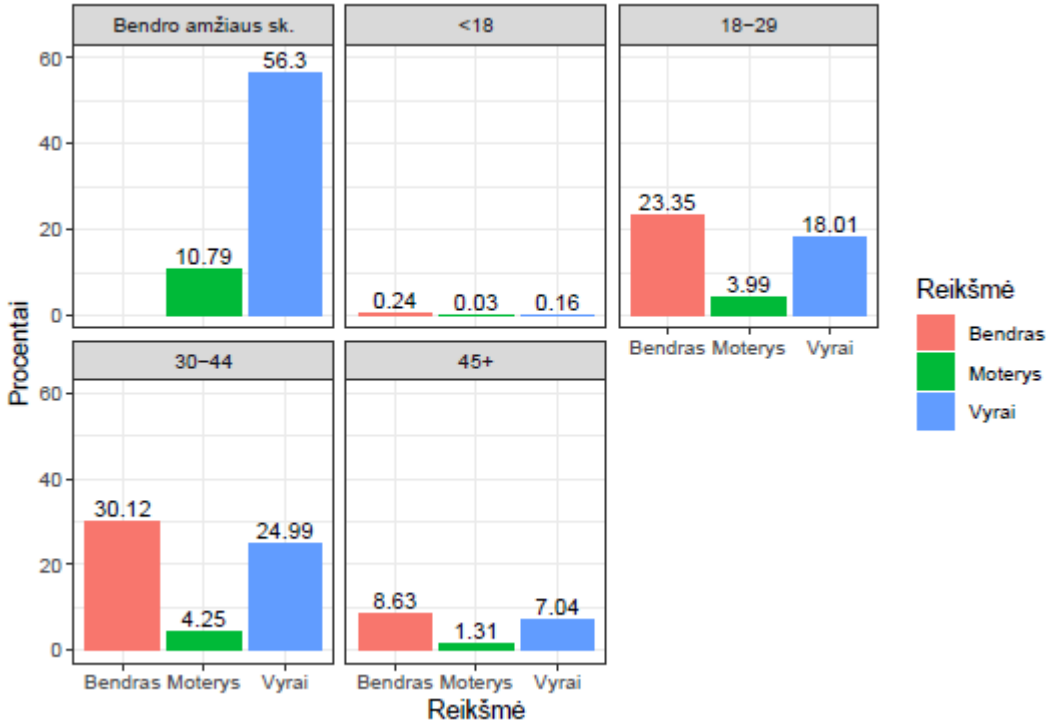
Reikia atsižvelgti ir į tai, kad dabartinės kainos (lentelėje pateiktos 2020-05-09 kainos) gali skirtis nuo kainų po filmų išleidimo datos. Pagrindinis tikslas - pasižiūrėti, kurio filmo kainos yra didesnės.

Pagal lentelėje pateiktus duomenis, galima teigti, kad mažesnias pajamas generuojančio filmo "Ready" DVD/Blue-ray diskų kainos didesnės, nei filmo "Beast" kainos.

Siekiant susidaryti bendrą vaizdą apie kino teatruose parduodamus bilietus, verta identifikuoti filmų auditoriją, nes bilietų kainos vaikams ir suaugusiam skiriasi.

Naudojami IMDB puslapio pateikti duomenys apie amžiaus ir lyties pasiskirstymą vertinant filmus (žr. pav.5).

5 pav.: Ready filmo auditorijos amžiaus pasiskirstymas



Galima daryti išvadą, kad vertinant abu filmus buvo aktyvesni vyrai (apytikliai sudarė 56 procentus visų vertinusių) ir didžiausią dalį, apytkilsiai 25proc., sudarė 30-44 amžiaus grupės vyrai. Taip pat aktyvi ir 18-29 amžiaus grupė, kuri "Ready" sudaro apytiklsiai 18proc., o "Beast" 20proc. vyrų.

Nagrinėjant pažeidimų mastą pasaulyje, pasiskirstymą tarp šalių, skirtingų lyčių, amžiaus grupių, ar koreliacijos tarp piratavimo ir pajamų lygio analizėje, bus naudojami tokie duomenys:

1. 2020 metų "Worldmeter" puslapio duomenys apie populiacijos dydį [12];
2. 2019 metų Euromonitor duomenys apie šalis - pajamų lygį, lyčių pasiskirstymą, amžiaus grupių pasiskirstymą;
3. 2019/07 - 2020/04 metų dieniniai UAB "Markmonitor" įmonės duomenys apie nelegalaus turinio platinimą, naudojant P2P techologiją, pažeidimų skaičių kiekvienoje pasaulio šalyje. Šis skaičius atspindi paspaudimų ant unikalaus kodo (file hash), kuriuo vartotojas dalinasi P2P technologija paremtuose puslapiuose;
4. Pasaulio Banko 2019 metų duomenys apie šalių pajamų grupes [13].

Siekiant palyginti šalis ir identifikuoti, kuriuose regionuose pažeidimų užfiksuota daugiausiai, naudoti šalių populiacijos duomenys užfiksuoti 2019 metais ir apskaičiuotas pažeidimų skaičius 100 tūskt. gyventojų.

Nagrinėjama 238 šalys, didžiausias pažeidimų skaičius (3 didžiausios reikšmės) 100 tūkst. gyventojų užfiksuotas Antarktidoje - 11018.52, Falklando salose - 9942.53, Bulgarijoje - 6787.35.

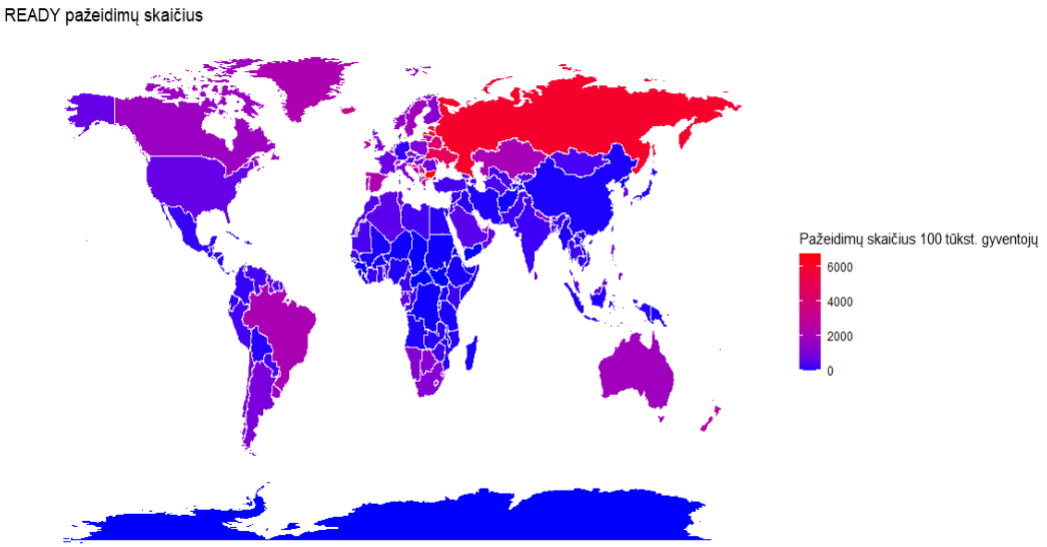
Antarktidoje laikotarpiu 2019/07-2020/04 fiksuoti 595 pažeidimai, tikėtina, kad mokslininkai atliekantys tyrimus, ar atkeliavę dirbti, kaip įgulos nariai, siunčiasi filmus nelegaliai.

Antarktidos regiono populiacija rasta šaltiniuose [14], kuriuose teigiama, kad gyventojų kiekis kinta priklausomai nuo sezono, žiemą apie 1100, o vasarą 5400 (2016 metų duomenimis), kurių pagrindinę dalį sudaro mokslininkai ir įgulos darbuotojai. Kadangi oficialios statistikos apie populiaciją nėra, analizėje naudojamas vasaros sezono lankytojų skaičius, nes nagrinėjamas laikotarpis atitinka vasaros sezoną Antarktidoje.

Antarktida turi stebėtinai daug pažeidimų, kai lankytojų - gyventojų skaičius siekia apie 5400 (imtas didžiausias galimas lankytojų - gyventojų skaičius vasaros sezonu, kuris ten tęsiasi nuo spalio iki vasario mėnesio), o pažeidimų užfiksuota 595, todėl bandoma pašalinti šį regioną, kaip išskirtį. Dar viena pašalinta išskirtis, kad geriau apžvelgti situaciją žemynuose, tai yra Folklandų (Malvinų) salos, kur populiacija: 3480, o pažeidimų skaičius 346, 100 tūkst. gyventojų tai sudaro net 9942.53 pažeidimus. Pasaulio pažeidimų žemėlapis pavaizduotas 6 paveikslėlyje.

Pašalinus išksirtis, daugiausiai pažeidimų 100 tūkst. gyventojų užfiksuota Bulgarijoje, Latvijoje, Juodkalnyjoje ir Rusijoje.

6 pav.: Ready pažeidimų skaičius



4.2 Demografinių ir ekonominių rodiklių apžvalga

Skyriuje trumpai aptariami pagrindiniai analizėje naudojami nepriklausomi kintamieji, kurie galėtų turėti poveikį pažeidimų skaičiui.

- **Pajamų grupė** - kokybinis kintamasis, 216 šalių, metinis 2019 m. rodiklis.

Potencialiems piratams identifikuoti, gali padėti sąryšio tarp šalies pajamų klasės įvertinimas. Naudojami Pasaulio Banko duomenys apie šalis ir joms priskiriamas pajamų grupes pagal 2019 metų vertinimo skalę (žr. lent. 3).

3 lentelė: Pajamų grupės

Grupė	**2019m. USD**	**Šalių skaičius**	**Ready Pažeidimų vidurkis**	**Ready std.**
Žemos pajamos	<1026	31	508.04	970.83
Žemesnės-vidutinės pajamos	1026 - 3995	46	1194.87	1658.05
Auskštesnės-vidutinės pajamos	3996 - 12375	59	3268.98	3758.68
Aukštos pajamos	>12375	805277.90	3821.69	

Lentelėje pateiktas šalių pasiskirstymas pagal priskiriamą pajamų grupę ir pagrindinės pažeidimų charakteristikos - vidurkis ir standartinis nuokrypis. Didžiausią dalį sudaro aukštų pajamų grupei priklausančios šalys - 37 procentai bei aukštesnių-vidutinių pajamų grupės šalys - 27.3 procento, atitinkamai didesnis ir pažeidimų vidurkis bei standartinis nuokrypis. Aukštai pajamų grupei priskirtose šalyse užfiksuota santykinai daugiausiai pažeidimų (žr. lent. 3).

- **Bendrosios nacionalinės pajamos** (GNI) **vienam gyventojui** - kiekybinis kintamasis, 186 šalims, metiniai 2019 m. duomenys.

Didžiausias reikšmes, kaip ir tikėtasi, įgyja Šveicarija, Norvegija, Islandija ir Liuksemburgas. Mažiausios reikšmės priskiriamos Afrikos žemyno šalims: Burundi, Malavis, Nigeris.

- **BVP vienam gyventojui** - kiekybinis kintamasis, 216 šalių, 2019 metų duomenys.

5 didžiausią reikšmę turinčios šalys aprašytos 4-oje lentelėje.

4 lentelė: BVP vienam gyventojui

Šalis	**BVP vienam gyenojui USD**
Monakas	183 tūkst.
Lichtenšteinas	182 tūkst.
Liuksemburgas	114 tūkst.
Bermuda	106 tūkst.
Kaimanų salos	91 tūkst.

- **Amžiaus grupių pasiskirstymas** - kiekybinis kintamasis, 191 šalis, 2019 metų duomenys.

  5 lentelėje pateiktos amžiaus grupės ir šalys. kuriose nagrinėjama grupė sudaro didžiausią procentą populiacijos.

5 lentelė: Amžiaus grupių pasiskirstymas šalyse

Amžiaus grupė	**Šalis**	**Procentas nuo populiacijos**
65+	Japonija	28.39%
45-49	Kuveitas	11.06%
40-44	Kuveitas	12.71%
35-39	Jungtiniai Arabų Emyratai	14.82%
30-34	Omanas	16.32%
25-29	Maldyvai	16.04%
20-24	Nepalas	10.84%
15-19	Centrinė Afrikos Respublika	12.06%

IMDB puslapio duomenimis buvo nustatyta, kad aktyviausiai filmais domėjosi 30-44 amžiaus grupės vyrai, todėl naudosime amžiaus grupę, kaip papildomą regresorių logistiniame modelyje, amžiaus grupių poveikiui pažeidimams identifikuoti.

- **Lyties pasiskirstymas procentais** - kiekybinis kintamasis, 216 šalių, 2019 metų duomenys.

Lyčių pasiskirstymas šalyse pakankamai vienodas ir svyruoja santykiu 40/60 vyrų ir moterų.

5 Analizė

5.1 Koreliacinė analizė

Skyriuje trumpai nagrinėjamos koreliacijos, tikrinamos hipotezės ryšiui tarp kintamųjų nustatyti. Skaičiuojamas Spirmeno (angl. Spearman) koreliacijos koeficientas, nes reikalauja mažiau apribojimų nei skaičiuojant Pirsono koreliacijos koeficientą - kintamųjų skirstiniai neprivalo tenkinti normalumo sąlygos, tikrinamas nebūtinai tiesiškas ryšys. Hipotezės:

$$\begin{cases} H_0 : \text{koreliacijos koeficientas lygus } 0 \\ H_1 : \text{koreliacijos koeficientas nelygus } 0 \end{cases}$$

Ryšys nustatinėjamas tarp kintamųjų:

- BVP vienam gyventojui ir pažeidimų skaičiaus;
- vyrų procento šalyse ir pažeidimų skaičiaus.

Rezultatai pateikti 6 lentelėje.

6 lentelė: Spirmeno koreliacijos testas

Spirmeno koreliacija	ρ_{x_i}	p-reikšmė
BVP vienam gyventojui	0.73	<0.05
Vyrų lyties procentas	0.05	0.52

Rezultatai rodo, kad yra statistiškai reikšmingas ryšys tarp kintamųjų BVP vienam gyventojui ir filmo Ready pažeidimų skaičiaus, nes p-reikšmė mažesnė už pasirinką reikšmingumo lygmenį 0.05. Kadangi $\rho_{x_i} > 0$, tai reiškia, kad didėjant pažeidimų skaičiui, didėja ir BVP vienam gyventojui reikšmė. Nenustatytas statistiškai reikšmingas ryšys tarp vyrų procento populiacijoje ir pažeidimų skaičiaus šalyse. Tačiau analizėje šį procentą naudosime, kad būtų galima daryti aiškesnį pirato "portretą".

Būtų įdomu panagrinėti duomenis apie gyventojų lankymosi kino teatre tendencijas, tačiau tokių, viešai prieinamų duomenų - nėra. Bandyta naudotis Euromonitor pateiktais 2019 metų duomenimis, tačiau stebėjimai fiksuoti tik 80 šalių, todėl šio kintamojo tolimesnėje analizėje nenaudosime. Tačiau pabandyti patikrinti Pirsono koreliacijos hipotezes - galima. Tiriamas ryšys tarp Ready filmo pažeidimų skaičiaus 100 tūkst. gyventojų ir apsilankymo kino teatre vienam gyventojui kartų. Gautas rezultatas, kad kintamieji susiję, nes p-reikšmė Spirmano teste lygi $0.018 < 0.05$ t.y. koreliacijos koeficientas nelygus 0.

5.2 Pajamų grupės ir pažeidimų priklausomybės analizė

Skyriuje aprašoma dispersinė analizė, kurios tikslas įvertinti priklausomybės stiprumą ir nustatyti, kurių pajamų grupių vidurkiai statistiškai reikšmingai skiriasi. Ar gyventojų didesnės pajamos lemia mažesnį audiovizualinių kūrinių teisių pažeidimų internete skaičių.

5.2.1 Pajamų grupių vidurkių palyginimas

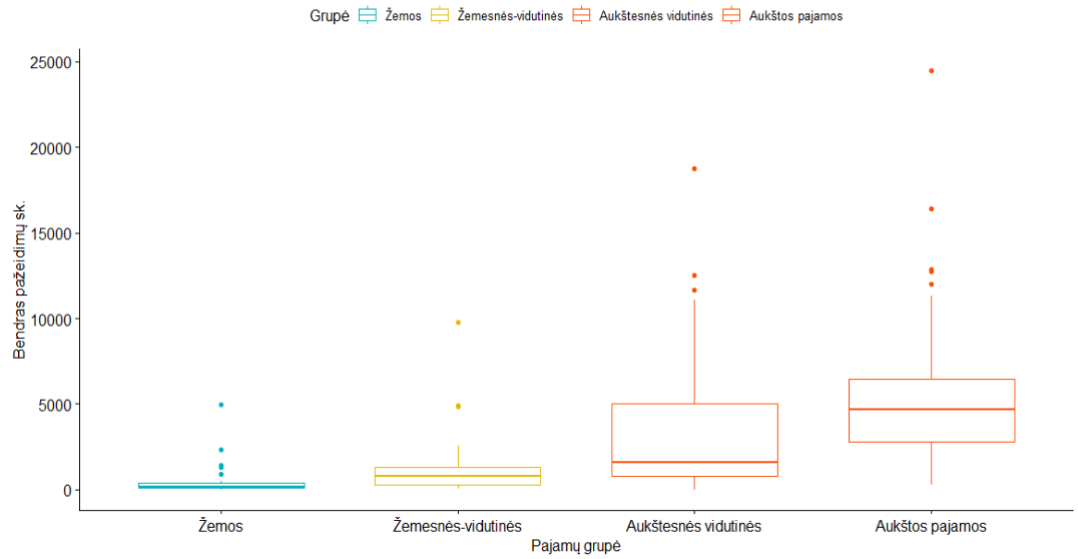
Taikoma **Welch ANOVA** bendro filmų Ready ir Beast pažeidimų skaičiui šalyse (laikotarpio 2019-07-01 - 2020-04-30) ir pajamų grupės į kurią šalys patenka, priklausomybei tirti.

Pirmiausia atliekama duomenų analizė išskirtims, normalumui ir dispersijai patikrinti.

- Išskirčių apžvalga:

7 paveikslėlyje pateikiamas grafikas, kuris vaizduoja duomenų charakteristiką kiekvienoje grupėje. Matoma keletas išskirčių. Tolimesnėje analizėje bus naudojami transformuoti (logaritmuoti) duomenys ir duomenys pašalinus didžiausias išskirtis (t.y. labiausiai nutolusius stebėjimus, pvz.: tarp 1 ir 3 kvantilio), vėliau rezultatai yra lyginami nustatyti iškirčių poveikį analizei.

7 pav.: Išskirtys



- Normalumo prielaidos tikrinimas - Šapiro - Vilko testas (angl. Shapiro–Wilk test)

$$\begin{cases} H_0 : \text{Imtis normaliai pasiskirsčiusi} \\ H_1 : \text{Imtis nėra normaliai pasiskirsčiusi} \end{cases}$$

Testo rezultatai pateikiami 7 lentelėje. Matoma, kad tik pusė pajamų grupių tenkina normalumo prielaidą. Aukštesnių - vidutinių ir aukštų pajamų grupės nėra vienodai pasiskirsčiusios. Tačiau atliekant testą duomenims su pašalintom iškirtim H_0 hipotezė priimama visoms grupėms, nes p-reikšmė < 0.05 pasirinktą reikšmingumo lygmenį.

7 lentelė: Šapiro - Vilko testo rezultatai

Šapiro - Vilko testas	**p-reikšmė**	**Statistika**	**p-reikšmė be išskirčių**	**Statistika**
Žemos pajamos	0.13	0.95	0.7	0.92
Žemesnės - vidutinės pajamos	0.49	0.98	0.49	0.92
Aukštesnės - vidutinės pajamos	<0.05	0.82	0.44	0.83
Aukštos pajamos	0.01	0.96	0.05	0.48

- Homogeniškumo tikrinimas - Levene testas.

$$\begin{cases} H_0 : \textit{duomenys homogeniški} \\ H_1 : \textit{duomenys nėra homogeniški} \end{cases}$$

Kad tenkinti normalumo prielaidą, duomenys logaritmuojami. Tikrinamas grupių dispersijų homogeniškumas. Pagal lentelėje pateiktus duomenis (žr. 8 lentelę) hipotezė apie skirtingų grupių dispersijų lygybę atetama abiem atvejais, nes p-reikšmės mažesnės už pasirinktą reikšmingumo lygmenį 0.05. Teigiame, kad duomenys nėra homogeniški, tai reiškia, kad skirtumas tarp grupių dispersijų yra statistiškai reikšmingas.

8 lentelė: Levene testo rezultatai

Levene testas	**p-reikšmė**	**Levene statistika**	**Laisvės laipsnis**
Logaritmuoti duomenys	<0.05	6.71	3
Logaritmuoti duomenys be išskirčių	<0.05	10.1	3

Atsižvelgiant į gautus rezultatus ir nustačius statistiškai reikšmingą skirtumą tarp grupių dispersijų, pasirenkamas Welch ANOVA modelis nehomogeniškoms imtims. Rezultatai pateikti 9 lentelėje. Kadangi p-reikšmė < 0.05, atmetame H_0 ir teigiame, kad yra statistiškai reikšmingas skirtumas tarp pajamų grupių vidurkių. Taikant keletą būdų nustatyti pajamų grupės ir pažeidimų skaičiaus sąryšiui daromos išvados - vienodos.

9 lentelė: Welch ANOVA rezultatai

Levene testas	**p-reikšmė**	**F statistika**	**Laisvės laipsnis**
Logaritmuoti duomenys	<0.05	59.43	3
Logaritmuoti duomenys be išskirčių	<0.05	86.54	3

Kadangi Welch ANOVA testas nenusako, tarp kurių pajamų grupių yra statistiškai reikšmingi vidurkių skirtumai, naudosime **Games Howel Post-Hoc kriterijų** duomenims su nevienodomis dispersijomis. Rezultatai pateikti 10 lentelėje (paryškinti duomenys aprašo duomenis be išskirčių). Tikrinama hipotezė:

$$\begin{cases} H_0 : \mu_i - \mu_j = 0 \\ H_1 : \mu_i - \mu_j \neq 0 \end{cases}$$

10 lentelė: Game-Howell testo rezultatai

Game Howell testas	**t-statistika**	**laisvės laipsnis**	**p-reikšmė**	**t-statistika**	**laisvės laipsnis**	**p-reikšmė**
Žemesnės-vidutinės-Žemos	3.96	47.36	<0.05	4.11	56.50	<0.05
Aukštesnės vidutinės-Žemos	5.89	60	<0.05	7.89	57.52	<0.05
Aukštos pajamos-Žemos	10.10	34.73	<0.05	13.54	32.35	<0.05
Aukštesnės vidutinės-Žemesnės-vidutinės	3.05	100.95	0.01	4.25	99.94	<0.05
Aukštos pajamos-Žemesnės-vidutinės	9.85	69.22	<0.05	10.41	61.94	<0.05
Aukštos pajamos-Aukštesnės vidutinės	4.28	75.81	<0.05	5.06	79.07	<0.05

Išvada: Priimama alternatyvi hipotezė, teigiama, kad skirtumas yra statistiškai reikšmingas tarp visų pajamų grupių vidurkių.

5.2.2 Pajamų grupių medianos palyginimas

Palyginti ir papildyti analizę naudosime neparametrinį ANOVA analogą, Kruskal-Wallis testą, ištirti grupių skirstinių vienodumą. Tikrinama hipotezė:

$$\begin{cases} H_0 : F_1(x) = F_2(x) = ... = F_k(x) \\ H_1 : \exists \text{ bent viena nelygybė} \end{cases}$$

Testo rezultatų lentelė 11 patvirtina, kad yra statistiškai reikšmigas skirtumas tarp pajamų grupių skirstinių.

11 lentelė: Kruskal-Wallis testo rezultatai

Kruskal Wallis testas	**Chi-statistika**	**laisvės laipsnis**	**p-reikšmė**
Duomenys	100.69	3	<0.05
Duomenys be išskirčių	103.57	3	<0.05

Nors ir Kruskal-Wallis atveju Post-Hoc kriterijai dažniausiai netaikomi, tačiau yra būdas tai padaryti ir papildyti analizę Dunn testu apie pajamų grupių vidurkių skirtumus. Rezultatai pateikti 12 lentelėje, nustatyti statistiškai reikšmingi skirumai tarp grupių vidurkių.

12 lentelė: Dunn testo rezultatai

Dunn testas	**Z-statistika**	**p-reikšmė**	**Z-statistika**	**p-reikšmė**
Žemesnės-vidutinės-Žemos	-2.28	<0.05	-2.26	<0.05
Aukštesnės vidutinės-Žemos	-5.36	<0.05	-5.39	<0.05
Aukštos pajamos-Žemos	8.86	<0.05	8.78	<0.05
Aukštesnės vidutinės-Žemesnės-vidutinės	-3.36	<0.05	-3.58	<0.05
Aukštos pajamos-Žemesnės-vidutinės	7.27	<0.05	7.6	<0.05
Aukštos pajamos-Aukštesnės vidutinės	3.99	<0.05	4.07	<0.05

5.2.3 Rezultatai

Šalių pajamų grupių pažeidimų vidurkiai ir medianos statistiškai reikšmingai skiriasi. Naudojantis Post-Hoc kriterijais nustatyta, kad statistiškai reikšmingas skirtumas yra tarp visų pajamų grupių.

5.3 Klasterinė laiko eilučių analizė

Siekiant nustatyti pažeidimų skaičiaus kitimo tendencijas pasaulyje, tinkamai paruošti šalinimo technologijas pažeidimų pikams ateityje ir parengti regioninę autorių teisių apsaugos

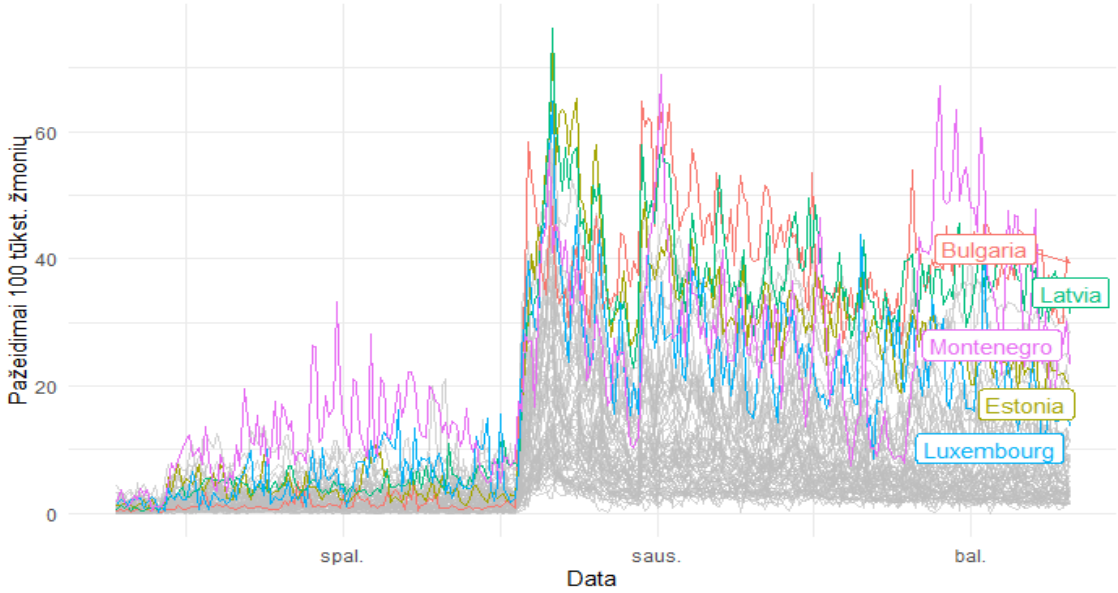
užtikrinimo strategiją, svarbu identifikuoti daugiausiai pažeidimų turinčius regionus, atrasti sąryšius tarp jų. Vienas tinkamiausių būdų tai padaryti - klasterinė laiko eilučių analizė.

Klasterinėje laiko eilučių analizėje bus naudojami dieniniai Ready filmo pažeidimų 100 tūkst. gyventojų duomenys laikotarpiu **2019-07-01 - 2020-04-30** (Beast filmo analizės rezultatai bus pateikti priede 8.1).

Analizei atrenkamos "reikšmingos" šalys. Kadangi klientai dažniausiai koncetruojasi į šalis, kuriose piko metu užfiksuojama daugiausiai pažeidimų, pašalinamos šalys, kuriose pažeidimų skaičius neužfiksuotas (na) ir šalys, kurių nagrinėjamo laikotarpio pažeidimų skaičius mažesnis už bendrą filmo Ready pažeidimų skaičiaus medianą. Pagrindins tikslas - šalių skaičiaus mažinimas, pavojingiausių šalių išskyrimas, kad nustatyti klasteriai būtų kuo informatyvesni.

Grafike 8 matomi neapdoroti duomenys ir paryškintos daugiausiai pažeidimų turinčios šalys - Bulgarija, Latvija, Estija, Liuksemburgas ir Juodkalnija.

8 pav.: Neapdorotos Ready filmo laiko eilutės



Prieš pradedant analizę, duomenys yra **standartizuojami**, kad būtų adekvačiai lyginamos eilutės: $\frac{x_i - \mu}{\sigma}$, kur i - šalies indeksas, $X = (x_1, x_2, ..., x_n)$ - šalies laiko eilutė, μ - eilutės X vidurkis, σ - standartinis nuokrypis, $\mu = \frac{\sum_{i=1}^{n} x_i}{n}$, $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$.

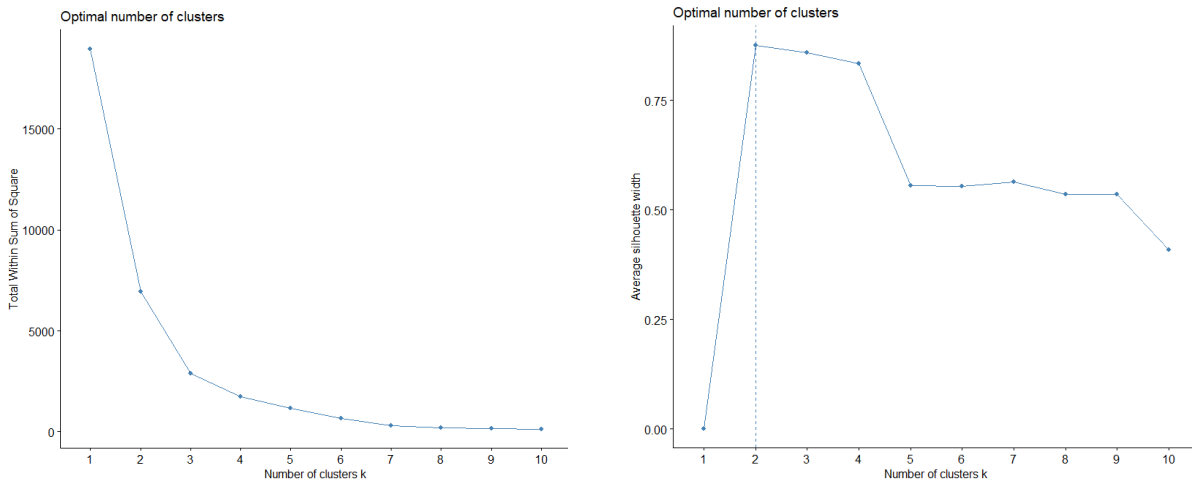
Ready filmo klasterinėje analizėje naudojamas **Euklido atstumo** matas, nustatyti mažiausią atstumą tarp laiko eilučių (Beast filmo analizė atlikta naudojantis Canberra metodu, kad būtų galima palygininti rezultatus). Parenkamas geriausias jungimo metodas (AGNES), duomenims taikant keturis pagrindinius jungimo tipus, Parinkimo taisyklė - didžiausia apskaičiuota reikšmė (žr. lent. 13).

13 lentelė: Jungimo metodo parinkimo rezultatai

	Klasterio centrų jungimo metodas	**Artimiausio jungimo metodas**	**Tolimiausio kaimyno metodas**	**Ward metodas**
Klasterizavimo struktūros stiprumo matas	0.86	0.80	0.91	0.96

Rezultatai rodo, kad tinkamiausias metodas bus **Ward metodas**, nes jis identifikuojamas, kaip turintis stipriausią klasterizavimo struktūrą.

Ieškant optimaliausio klasterių skaičiaus, taikomas Elbow ir Silhouette metodas, kurių rezultatai atitinkamai vaizduojami grafiškai 9 paveikslėlyje. Optimaliausias yra tas, kuris Elbow sudaro įgaubimą, o Silhouette funkcijos globalus maksimumas.

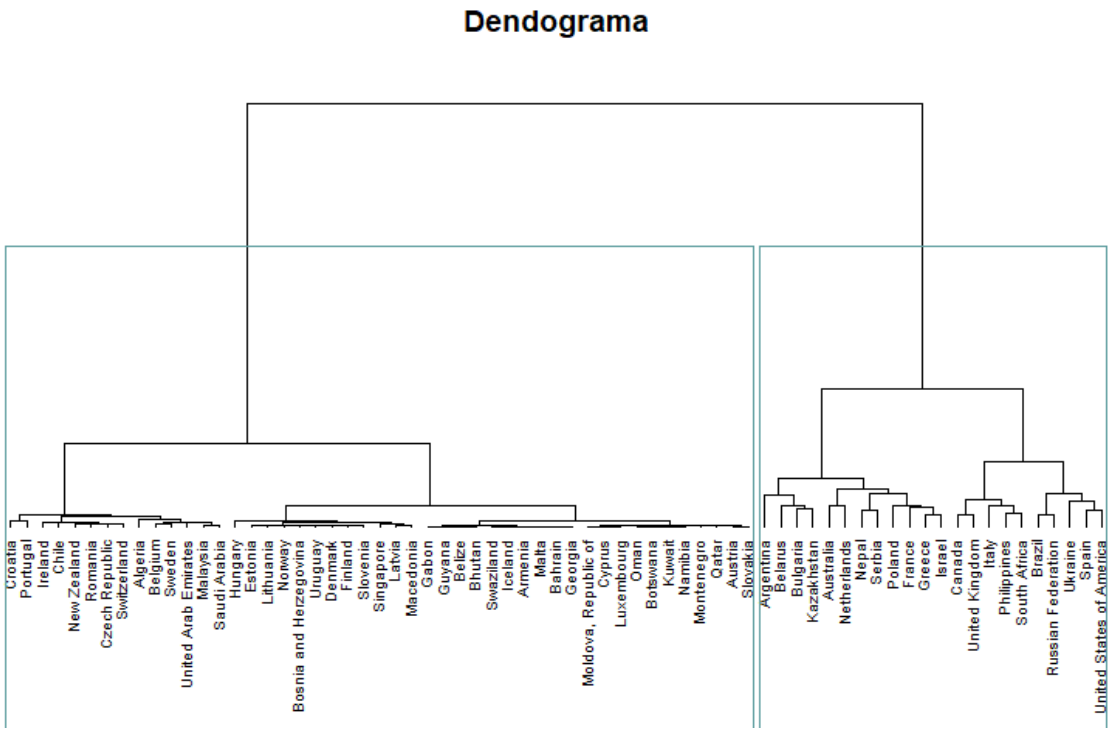


(a) Elbow (b) Silhouette

9 pav.: Elbow ir Silhouette metodai

Pasirinkus atstumo matą - atstumo matricai skaičiuoti ir tinkamiausią skaičiavimo metodą, brėžiama dendograma, kuri vaizduoja hierarchinį klasterizavimo metodą ir šalis, suskirstytas į pasirinktą klasterių skaičių. 10 grafikas pateikiamas atveju, kai pasirinktas klasterių skaičius lygus 2, tačiau analizuojant duomenis, buvo išbandyta keletas galimų klasterių skaičiaus kombinacijų.

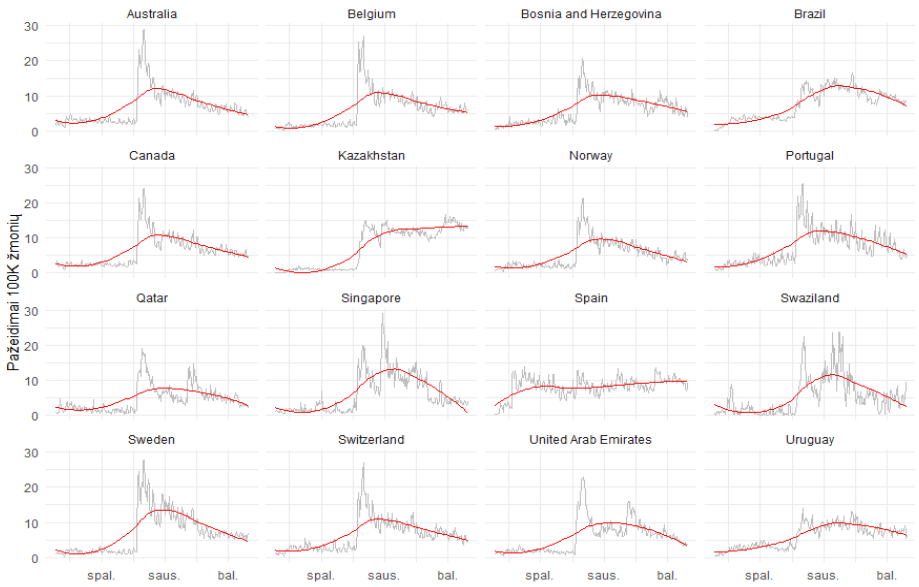
10 pav.: Dendograma



5.3.1 Rezultatai

Dendogramoje, kurios pjūvis suskirsto šalis į du klasterius (pirmąją grupę sudaro 47, antrąją - 22 šalys), antrojo klasterio charakteristika pateikta paveikslėlyje 11.

11 pav.: Šalys, patenkačios į antrąjį klasterį



Požymių lentelėje 14 pateiktos išbandytos klasterių skaičiaus reikšmės, atstumo skaičia-

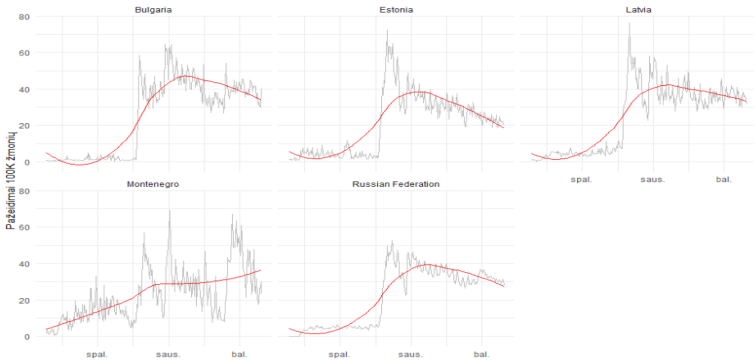
vimo metodai ir pastebėti požymiai.

14 lentelė: Požymių lentelė

Filmo pavadinimas	**Atstumo matas**	**Klasterių sk.**	**Požymis**
Ready	Euklido	2	Laiko eilučių tendencingas judėjimas: pakilimas ir nuosmukis
Ready	Canberra	5	Kaimyninių šalių regionai
Beast	Euklido	8	Politinis sąryšis
Beast	Canberra	8	Kalba

Pasinaudojus klasterine analize galima nustayti, kurios šalys elgiasi panašiai pvz.: Rusija, Latvija, Estija dažniausiai priskiriamos į vieną grupę, kaip tai matoma suskirsčius Beast filmo pažeidimų laiko eilutes į 8 klasterius, pasinaudojus Euklido atstumo matu (žr. pav. 12). Pastebima, kad autorių teisių pažeidimų tendencijos panašios kaimyninėse valstybėse, regionuose. Priežastis galėtų būti panaši informacinė aplinka, kultūra, geografinė padėtis, leidžianti naudotis kokybiškomis interneto paslaugomis.

12 pav.: Beast šalių pažeidimų klasteris



Matomas šalių pasiskirstymas pagal kalbas, t.y. atsiradus filmo kopijai su vertinmu tam tikra kalba, tikimasi užfiksuoti pažeidimų šuolį regionuose, kur ta kalba yra vartojama.

Remiantis klasterinės analizės duomenimis, galima prognozuoti pažeidimų kiekį po planuojamo filmo išleidimo datos regionuose.

5.4 Logistinė regresija

Sudarinėjama logistinė regresija pažeidimams 100 tūkst. gyventojų . Tikslas nustatyti priklausomybes tarp filmų pažeidimų internete ir ekonominių bei demografinių rodiklių. Pažeidimai skirstomi į grupes: šalys, kuriose jų užfiksuota daugiausiai ir šalys, kurios laikomos

mažiau pavojingomis. Priklausomąjį kintamąjį perkonstruojam į dvireikšmį, skiriam pagal medianą į 2 grupes (daug pažeidimų - 1, mažai - 0). Reikia paminėti, kad matematinis modelis bus sudaromas ne priklausomam kintamajam Y - pažeidimams, bet jo tikimybių santykio logaritmui (logit funkcijai).

Atliekama transformacija rodikliui BVP vienan gyventojui, nes nustatyta dešinioji asimetrija. **Transformuojama** pagal: $ln(\frac{1}{2} + x)$ taisyklę.

Atsižvelgiant į klasterinės analizės išvadas, įtraukiamas **regiono** kintamasis ir parenkamas geriausais modelis pagal AIC kriterijų.

5.4.1 Rezultatai

Pirmasis modelis konstruojamas įtraukiant 3 kintamuosius, atmetus nereikšmingas amžiaus grupes: BVP vienam gyventojui, vyrų procentą ir 30-34 amžiaus grupę. Pimojo modelio gauti rezultatai aprašomi 15 lentelėje.

$$ln(\frac{\widehat{p_1}}{1-\widehat{p_1}}) = -9.692 + 1.905 \cdot \texttt{BVP\_vienam\_gyventojui} - 21.200 \cdot \texttt{vyrai\_proc} + 46.668 \cdot \texttt{30-34},$$

čia $\widehat{p_1}$ - pirmojo modelio tikimybė rezultatui 1 (1 - daug pažeidimų).

15 lentelė: Pirmojo modelio regresorių aprašomoji lentelė

	Įverčiai	**Stand. paklaidos**	**z-reikšmė**	**p-reikšmė**
`Laisvasis narys`	-9.692	3.953	-2.452	0.0142
`BVP_vienam_gyventojui`	1.905	0.270	7.056	<0.05
`30-34`	46.668	20.840	2.239	0.0251
`vyrai_proc`	-21.200	9.782	-2.167	0.0302

Modelyje matosi, kad amžiaus grupė (30-34), BVP vienam gyventojui ir vyrų procentas šalyje yra reikšmingi kintamieji. Šis rezultatas patvirtina išvadas, padarytas nagrinėjant IMDB puslapio duomenis apie amžiaus grupes ir lyties pasiskirstymą vertinant filmą.

Atsižvelgus į klasterinės analizės išvadas, sudarinėjamas antrasis modelis ir įtraukiamas naujas kintamasis - regionas. Pagal suformuotus požymius klasterinėje analizėje (pvz: kalba) ir turint pakankamai duomenų apie rinkmenos charakteristiką (pažeidimas padarytas spaudžiant ant unikalaus kodo rinkmenai parsisiųsti tam tikra kalba), būtų galima konstruoti pseudokintamąjį. Šiuo atveju to nedarysime, nes ženkliai sumažinamas duomenų rinkinys ir rezultatai gaunami netikslūs.

$$ln(\frac{\widehat{p_2}}{1-\widehat{p_2}}) = -6.8013 + 0.00009 \cdot \texttt{BVP\_vienam\_gyventojui} + 1.6447 \cdot \texttt{Europos\_regionas} + 0.955 \cdot \texttt{30-34},$$

16 lentelė: Antrojo modelio regresorių aprašomoji lentelė

	Įverčiai	**Stand. paklaidos**	**z-reikšmė**	**p-reikšmė**
`Laisvasis narys`	-6.8013	4.204	-1.618	0.106
`BVP_vienam_gyventojui`	0.00009	0.00003	3.177	0.001
`Europos_regionas`	1.6447	0.213	1.680	<0.05
`30-34`	0.955	0.979	3.465	0.093

čia $\widehat{p_2}$ - antrojo modelio tikimybė rezultatui 1 (1 - daug pažeidimų).

16 lentelėje matoma įtraukus regioną, kaip papildomą nepriklausomą kintamąjį, vyrų procentas pasidarė nereikšmingas. Pastebėta, kad Europos regionas - reikšmingas kitntamasis, kas reiškia, kad didžioji dalis filmo Ready pažeidimų užfiksuota Europos regione. Verta paminėti, kad Beast filmo logistinėje regresijoje užfiksuotas Azijos regiono statistinis reikšmingumas.

5.4.2 Modelio patikimumo tikrinimas

Nors modelis sudarytas iš mažai regresorių ir dėl duomenų trūkumo nėra adekvatus, tačiau pabandyta parinkti geresnį modelį ir tikrinti patikimumą. Geresnis modelis renkamas pagal AIC kriterijų bei deviacijos ir laisvės lapsnių santykį (žr. lent. 17).

17 lentelė: Modelių palyginimas

	AIC	**deviacihos ir df santykis**
Pirmasis modelis	139.92	0.71
Antrasis modelis	158.36	0.65

Pirmojo modelio informacinis indeksas AIC yra mažesni, kas reiškia, kad modelis yra geresnis bei deviacijos ir laisvės laipsnių santykis yra arčiau 1, kas reiškia, kad modelis gerai paaiškina duomenis.

Toliau nagrinėsime pirmąjį modelį. Sudarinėjame klasifikavimo lentelę (žr. lent.18), duomenis suskirstome į 2 grupes (80/20) į apmokamąją ir testo grupę, kad patikrinti, kaip gerai modelis prognozuoja. Didžioji dalis prognozių pasitvirtino (83 proc). Modelis Ready prognozuoja patikimiau, nei filmo Beast modelis (0.76 proc.).

Kintamųjų **interpretacija**: kad modelio rezultatai įgautų prasmę, skaičiuojami galimybių santykiai: e^{β_i} ir kintamasis - BVP vienam gyventojui, kuriam buvo pritaikyta transformacija apskaičiuojamas taip: $e^{e_i^{\beta}} - \frac{1}{2}$. Pavyzdžiui, pirmojo modelio BVP koeficiento reikšmė lygi 1.905, galimybių santykis bus 6.719, kas reiškia, kad BVP vienam gyventojui padidėjus vienetu, pažeidimų skaičius padidėja 6,72 karto, o pavertus atgal į USD gauname, kad BVP vienam gyventojui lygi $e^{e^{1.905}} - \frac{1}{2} = 827.82$, padidina pažeidimų mastą 6,72 karto.

18 lentelė: Klasifikavimo lentelė

	Prognozuota reikšmė	
Tikroji reikšmė	**0**	**1**
0	63	5
1	19	59

5.4.3 Išvados

Ready filmo logistinėje regresijoje nustatytas reikšmingas Europos regionas, o Beast filmo logistinės regresijos rezultatai rodo pažeidimų ir Azijos regiono sąryšį. Tai gali reikšti, kad skirtingo žanro filmai populiaresni skirtinguose regionuose. To priežastis - skirtingos kultūros, jų vertybės.

Stipriausias sąryšis pastebėtas tarp BVP vienam gyventojui ir pažeidimo masto pasaulyje. Reikšmingas kintamasis ir vyrų procentas šalies populiacijoje bei vidutinė amžiaus grupė. Galima daryti išvadą, kad vyresnės kartos atstovai linkę naudotis senesnėmis technologijomis ir platformomis dėl paprastumo ir novatoriškumo vengimo. Nereikšmingas kintamasis buvo 20-30 metų amžiaus grupė, galima sakyti, kad jaunesnė karta linkus naudoti naujesnėmis technologijomis (Netflix, Apple TV). Taip pat bandytas skaičiuoti pažeidimų sąryšis su interneto greičiu, bet šis kintamasis neįtraukiamas į bendrą modelį, nes interneto greitis labiausiai priklauso nuo geografinės padėties, todėl nustatyta sąsaja su tam tikra šalimi, gali būti neteisinga.

Bandytas tikrinti ir sezoniškumas (grafikai priede 18), nes abiejų filmų laiko eilučių grafikuose pastebėtas didelis pakilimas gruodžio-sausio mėnesį. Tai galėtų būti paaiškinta šventiniu laikotarpiu, kai gyventojai turi daugiau laiko ir leidžia jį su šeima, ar draugais.

6 Išvados bei rekomendacijos

1. Anlaizėje buvo naudojami dviejų filmų (Beast ir Ready) pažeidimų duomenys užfiksuoti P2P tinkle. Remiantis jais atlikta:

   - Dispersinė analizė - ištirta priklausomybė tarp šalims priskirtos pajamų grupės ir pažeidimų;
   - Klasterinė laiko eilučių analizė - nustatytos pažeidimų tendencijos ir požymiai;
   - Sukonstruoti logistinės regresijos modeliai - nustatyti sąryšiai tarp pažeidimų ir ekonominių bei demografinių rodiklių.

2. Atlikus koreliacinę analizę pažeidimų sąryšiui su BVP vienam gyventojui rodikliu ir vyrų procentu populiacijoje įvertinti, gauti tokie rezultatai:

   - tarp pažeidimų ir BVP vienam gyventojui nustatyta teigiama, statistiškai reikšminga koreliacija ($\rho = 0.73$). BVP vienam gyventojui augant, didėja pažeidimų skaičius, todėl ekonomiškai stipresnėse šalyse tikėtina užfiksuoti daugiau pažeidimų.
   - tarp pažeidimų ir vyrų procento populiacijoje nustatyta teigiama, bet statistiškai nereikšimnga koreliacija. ($\rho = 0.05$). Vyrų procentas šalies populiacijoje nenulemia pažeidimų šalyje skaičiaus didėjimo. Konstruojant modelį pažeidimų prognozėms nustatyti, toks rodiklis būtų svarstytinas.

3. Dispersinė analizė parodė, kad yra statistiškai reikšmingas skirtumas tarp visų keturių (aukštų, aukštų-vidutinių, vidutinių, žemų-vidutinių, žemų) pajamų grupių vidurkių. Vidutiniškai daugiausiai pažeidimų užfiksuota aukštai pajamų grupei priklausančiose šalyse. Piratavimas labiau paplitęs didesnes pajamas generuojančiose regionuose.

4. Remiantis klasterinės analizė rezultatais nustatyti pagrindiniai 4 šalių susiskirstymo požymiai:

   - laiko eilučių tendencingas judėjimas;
   - kaimyninių šalių regionai;
   - politinis sąryšis;
   - vartojama kalba.

   Užfiksavus pirmąją kopiją regione, kuris patenka į priskirtą grupę pagal požymį (pvz.: požymis - kaimyninių šalių regionai), ruošiamasi efektyviai šalinti kopijas ir į tą pačią grupę pagal požymį patenkančiose šalyse (pvz.: Latvijoje nustayta pirmoji kopija, todėl ruošiamasi pažeidimų šalinimui ir Estijoje).

5. Sukonstruoti 2 logistinės regresijos modeliai, parinktas geresnis pagal AIC kriterijų bei deviacijos ir laisvės laipsnio santykį. Sudaryta klasifikavimo lentelė, efektyviausio modelio patikimumas įvertintas 83 procentais. Reminatis logistinės regresijos rezultatais, galima teigti, kad stipriausia priklausomybė yra tarp ekonominio rodiklio - BVP vienam gyventojui ir pažeidimų skaičiaus, taip pat reikšmingas ir 30-34 amžiaus grupės procentas šalies populiacijoje, patvirtintas regioninis sąryšis, išskiriami Azijos ir Europos regionai. Prognozėms modelis nėra tinkamas dėl mažai jame esančių regresorių ir duomenų stokos.

6. Konstruojant modelį prognozėms, ateityje būtų naudinga įtraukti daugiau socialinių rodiklių, tokių kaip: išsilavinimas, nedarbas, nuosavybės teisės stiprumo indikatorius. Atlikti funkcinę duomenų analizę ir panagrinėti sezoniškumą. Įvertinti greitį nuo filmo išleidimo datos iki pirmos kopijos atsiradimo - šalinimo efektyvumui užtikrinti.

7 Literatūros sąrašas

1. Joost Poort, João Pedro Quintais, *Global Online Piracy Study*. Straipsnio internetinė prieiga:
   https://www.researchgate.net/publication/327026436_Global_Online_Piracy_Study

2. MUSO's 2017 Annual Piracy Reports, *Global Piracy Increases Throughout 2017, MUSO Reveals*. Straipsnio internetinė prieiga:
   https://www.muso.com/magazine/global-piracy-increases-throughout-2017-muso-reveals

3. IRDETO flipbook, *The Piracy Landscape has Web Video Replaced Peer to Peer*. Straipsnio internetinė prieiga:
   https://resources.irdeto.com/white-papers-e-books-reports/the-piracy-landscape-has-web-video-replaced-peer-to-peer

4. Čekanavicius V., Murauskas G., *Logistinė regresija*. Straipsnio internetinė prieiga:
   http://stat.vadoveliai.lt/files/LogRegSAS.pdf

5. Prof. habil. dr. Čekanavicius V., *INFERENCINĖ STATISTIKA SOCIALINIUOSE MOKSLUOSE*. Straipsnio internetinė prieiga:
   http://www.lidata.eu/index.php?file=files/mokymai/inferencine_2011/infer.html\&course_file=infer_3_3.html

6. Salvatore S.Mangiafico, *Kruskal–Wallis Test*. Straipsnio internetinė prieiga:
   https://rcompanion.org/rcompanion/d_06.html

7. Vilmantas Gėgžna, *Statistinės analizės pagrindai*. Straipsnio internetinė prieiga:
   https://mokymai.github.io/teorija-2020/ht-anova-ir-analogai.html

8. Hangcheng Liu, *Comparing W Comparing Welch's ANOVA, a Kruskal-W A, a Kruskal-Wallis test and tr allis test and traditional aditional ANOVA in case of Heterogeneity of Variance*. Straipsnio internetinė prieiga:
   https://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=5026&context=etd

9. Jon Starkweather, *Module 9: One-Way Analysis of Variance (ANOVA)*. Straipsnio internetinė prieiga:
   http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/ISSS_SC/Module009/isss_m91_onewayanova/node7.html

10. Eva Ostertagova, Oskar Ostertag, *Methodology and Application of the Kruskal-Wallis Test*. Straipsnio internetinė prieiga:

https://www.researchgate.net/publication/289442433_Methodology
_and_Application_of_the_Kruskal-Wallis_Test

11. Alexis Dinno, *Nonparametric pairwise multiple comparisons in independent groups using Dunn's test*. Straipsnio internetinė prieiga:
https://journals.sagepub.com/doi/pdf/10.1177/1536867X1501500117

12. Worldmeter data, *World Population: Past, Present, and Future*. Duomenų internetinė prieiga:
https://www.worldometers.info/world-population/

13. World Bank data, *GNI (current US)*. Duomenų internetinė prieiga:
https://data.worldbank.org/

14. World Population Review Article, *Antarctica Population 2020*. Duomenų internetinė prieiga:
https://worldpopulationreview.com/continents/antarctica-population/

15. AI Hub training article. *Training using a k-means algorithm with Tens or Flow*:
https://aihub.cloud.google.com/u/0/p/products%2F0e0d2ed0-5563-4639-b348-53a83ac4ff4e

16. Michel Kuntz, Julian Aliaj. *Clustering SVG Shapes*:
http://www.graphicalweb.org/2009/papers/5-Clustering_SVG_Shapes/

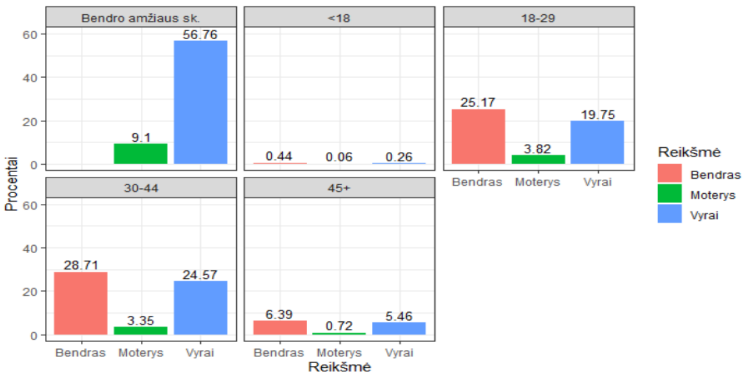
17. Creative Commons Attribution picture. *Type of Hierarchical Clustering*:
https://www.researchgate.net/figure/Overview-of-the-difference-between-agglomerative-and-divisive-hierarchical-clustering_fig1_304378246

18. Dex Lab analytics Article. *Hierarchical Clustering*:
https://www.dexlabanalytics.com/blog/hierarchical-clustering-foundational-concepts-and-example-of-agglomerative-clustering

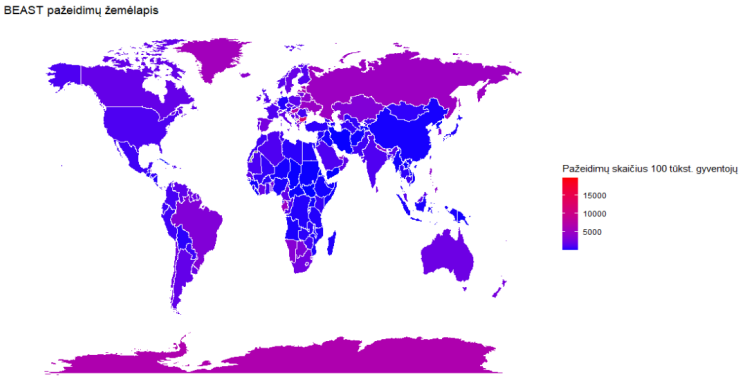
8 Priedai

8.1 Antrojo - Beast filmo analizės rezultatai

13 pav.: Beast filmo amžiaus pasiskirstymas



14 pav.: Beast filmo pažeidimų žemėlapis



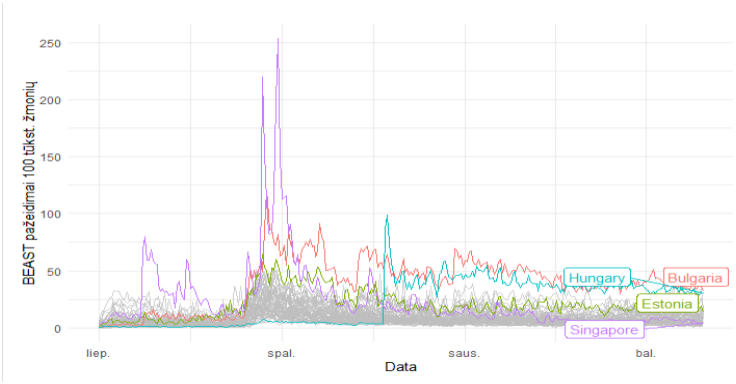
19 lentelė: Beast Pajamų grupės

Grupė	2019m. USD	Šalių skaičius	Ready Pažeidimų vidurkis	Ready std.
Žemos pajamos	<1026	31	414.96	784.36
Žemesnės-vidutinės pajamos	1026 - 3995	46	822.24	993.54
Auskštesnės-vidutinės pajamos	3996 - 12375	59	2190.94	2559.41
Aukštos pajamos	>12375	80	3333.41	2927.10

20 lentelė: Spirmeno koreliacijos testas Beast

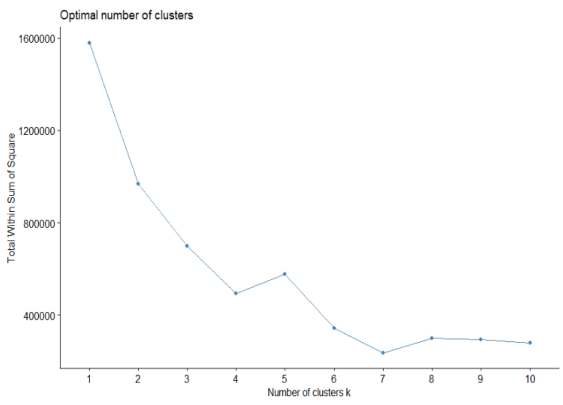
Spirmeno koreliacija	ρ_{x_i}	p-reikšmė
BVP vienam gyventojui	0.62	<0.05
Vyrų lyties procentas	0.06	0.39

15 pav.: Beast filmo žalios laiko eilutės

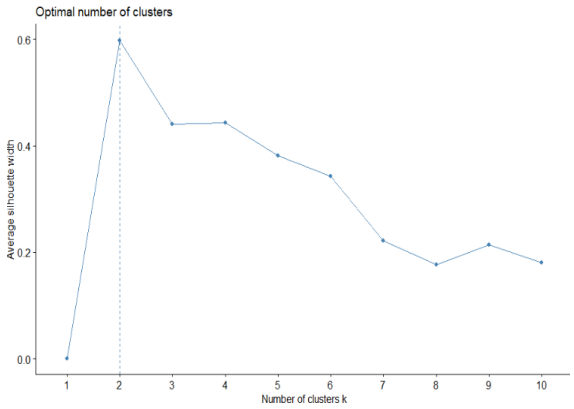


21 lentelė: Jungimo metodo parinkimo rezultatai Beast

	Klasterio centrų jungimo metodas	**Artimiausio jungimo metodas**	**Tolimiausio kaimyno metodas**	**Ward metodas**
Klasterizavimo struktūros stiprumo matas	0.88	0.86	0.90	0.93



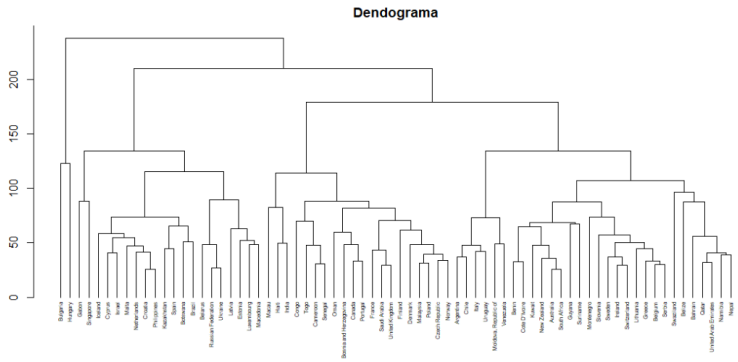
(a) Elbow Beast



(b) Silhouette Beast

16 pav.: Elbow ir Silhouette metodai Beast

17 pav.: Beast dendograma



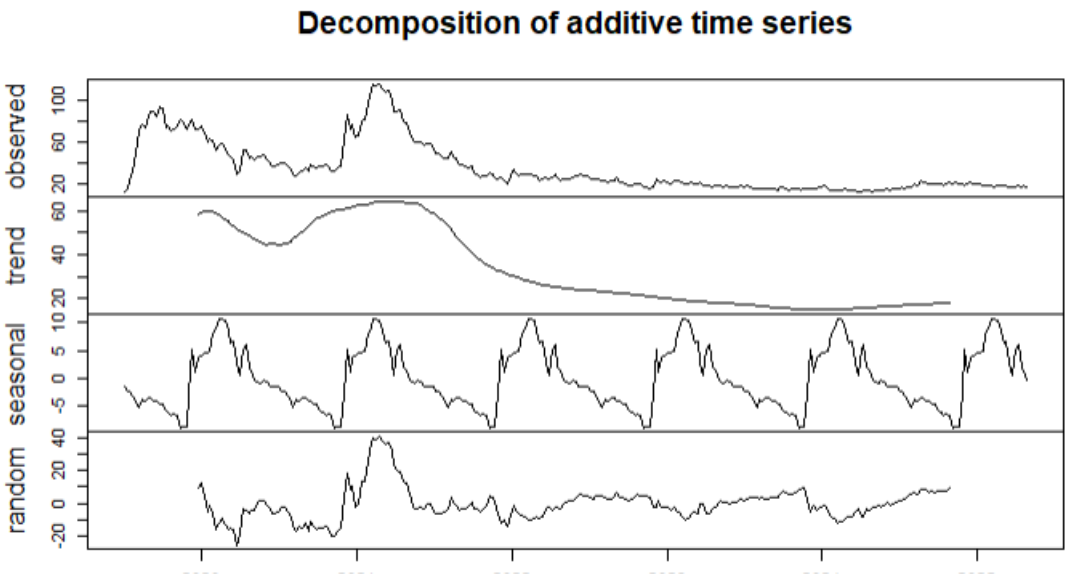
22 lentelė: Klasifikavimo lentelė Beast

	Prognozuota reikšmė	
Tikroji reikšmė	**0**	**1**
0	65	9
1	25	46

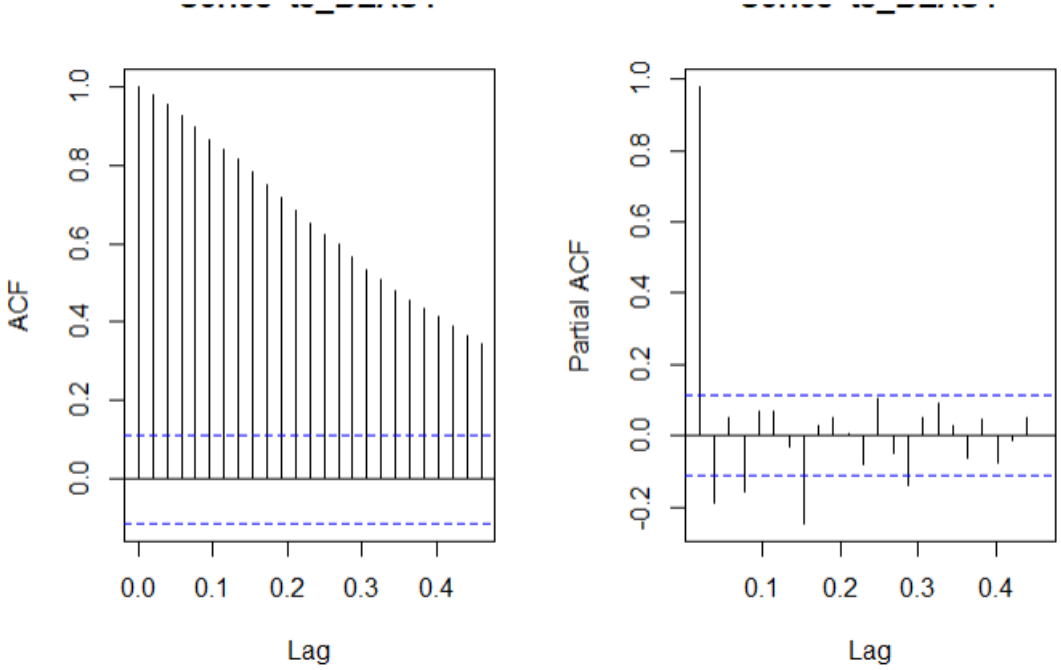
23 lentelė: Beast antrasis modelis

	Įverčiai	**Stand. paklaidos**	**z-reikšmė**	**p-reikšmė**
Laisvasis narys	-4.077	1.2437	0.381	0.00104
BVP_vienam_gyventojui	0.00007	0.000029	4.106	0.02095
Azijos regionas	0.592	-2.013	1.140	0.2542
40-44	0.9545	0.2422	1.900	0.01453

18 pav.: Sezoniškumas



19 pav.: ACF, PCF grafikai



8.2 R programos kodas

```
library(forecasst)
library(tseries)
library(writexl)
require(rugarch)
library(lubridate)
library(dplyr)
library(naniar)
library(dygraphs)
library(ggplot2)
library(lubridate)
library(magrittr)
library(rnaturalearthdata)
library(rgeos)
library(tmap)
library(sf)
theme_set(theme_bw())
library(data.table)
library(reshape2)
set.seed(1612070)
library(datasets)
e)
library(lemon)
library(kableExtra)
library(pander)
library(plyr)

#Filmų specifikacija:
kable(Filmai, format="latex", booktabs=TRUE,align = c("c", "c", "c", "c","c"))
%>% kable_styling(latex_options="scale_down",full_width = F,font_size = 7) %>%
column_spec(1:9,width = "5em")
kable(Filmai, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down")
kartai <- cbind(c(Filmai$‘Balsavusių žmonių skaičius‘[1]/
Filmai$‘Balsavusių žmonių
skaičius‘[2],
Filmai$‘Pajamos iš kino teatro pasaulyje USD‘[2]/
Filmai$‘Pajamos iš kino teatro pasaulyje USD‘[1],
Filmai$‘DVD pardavimai šalies viduje USD‘[2]/
```

```
Filmai$‘DVD pardavimai šalies viduje USD‘[1], Filmai$‘Blue-ray pardavimai
šalies viduje USD‘[2]/
Filmai$‘Blue-ray pardavimai šalies viduje USD‘[1],
Filmai$‘Biudžetas USD‘[2]/Filmai$‘Biudžetas USD‘[1]))
kartai <- t(kartai)
kartai <- as.data.frame(kartai)
rownames(kartai) <- "Santykis"
colnames(kartai) <- c("Balsavusių žmonių skaičius",
"Pajamos iš kino teatro pasaulyje USD", "DVD pardavimai šalies viduje USD",
"Blue-ray pardavimai šalies viduje USD", "Biudžetas USD")


#Filmų įvertinimo ir kainų santykis
kable(kartai, format="latex", booktabs=TRUE,
align = c("c", "c", "c", "c","c")) %>%
  kable_styling(latex_options="scale_down",
  full_width = F) %>% column_spec(1:5,width = "10em")
kainos <- matrix(c(16.12,20.73,8.66,12.39),ncol=2,byrow=TRUE)
colnames(kainos) <- c("DVD kaina","Blue-ray kaina")
rownames(kainos) <- c("Ready","Beast")
kainos <- as.table(kainos)
kable(kainos, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="striped",full_width = F)


#IMDB amžiaus ir lyties pasiskirstymas
#Ready
Ready_filmu_pasiskirstymas <- Ready_filmu_pasiskirstymas_1[-1]
rownames(Ready_filmu_pasiskirstymas) <- Ready_filmu_pasiskirstymas_1$...1
kable(Ready_filmu_pasiskirstymas, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down",full_width = F)
#Beast
Beast_filmu_pasiskirstymas <- Beast_filmu_pasiskirstymas_1[-1]
rownames(Beast_filmu_pasiskirstymas) <- Beast_filmu_pasiskirstymas_1$...1
kable(Beast_filmu_pasiskirstymas, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down",full_width = F)


#procentinis amžiaus ir lyties pasiskirstymas
```

```
procentiniai_Ready_filmu_pasiskirstymas <- round(Ready_filmu_pasiskirstymas/
458751*100,2)
procentiniai_Beast_filmu_pasiskirstymas <- round(Beast_filmu_pasiskirstymas/
287712*100,2)
data_1 <- melt(procentiniai_Ready_filmu_pasiskirstymas)
data_1$reiksme <- c("Bendras","Vyrai","Moterys")
ggplot(data_1, aes(reiksme, value,fill=reiksme)) + geom_bar(stat="identity")
+ facet_wrap(~ variable) + xlab("Reikšmė") + ylab("Procentai")+
labs(fill = "Reikšmė") + ylim(0, 60) + geom_text(aes(label=value),
vjust=-0.3, size=3.5)
fig <- function(width, heigth){
  options(repr.plot.width = width, repr.plot.height = heigth)
}
fig(10, 4)
#"Beast" filmas
data_2 <- melt(procentiniai_Beast_filmu_pasiskirstymas)
data_2$reiksme <- c("Bendras","Vyrai","Moterys")
ggplot(data_2, aes(reiksme, value,fill=reiksme)) + geom_bar(stat="identity")
+ facet_wrap(~ variable)+ xlab("Reikšmė") + ylab("Procentai")+
labs(fill = "Reikšmė") + ylim(0, 60) + geom_text(aes(label=value),
vjust=-0.3, size=3.5)
#Pažeidimų sk. 100 tūkst. gyventojų:
Mapsui_Ready <- data_1[,c(3,5)]
names(Mapsui_Ready) <- c("Country","Infringements")
#priskiriam reiksmes vieno salyno saloms
Mapsui_Ready$Infringements[89] <- 477+2139
Mapsui_Ready$Infringements[109] <- 477+2139
de<-data.frame("Antarctica",5400)
names(de)<-c("Salis","Populiacija")
populiacija <- rbind(populiacija,de)
populiacija$Salis <- dplyr::recode(populiacija$Salis
  ,'Brunei' = 'Brunei Darussalam',
 'Cabo Verde' = 'Cape Verde',
  'DR Congo' = 'Congo, The Democratic Republic of the',
  "Côte d'Ivoire" = "Cote D'Ivoire",
  'Curaçao' = "Curacao",
  "Czech Republic (Czechia)" = "Czech Republic",
  "Timor-Leste" = "East Timor",
  "Falkland Islands" = "Falkland Islands (Malvinas)",
```

```
  "Faeroe Islands" = "Faroe Islands",
  "Holy See" = "Holy See (Vatican City State)",
  "Iran" = "Iran, Islamic Republic of",
  "South Korea" = "Korea, Republic of",
  "North Korea" = "Korea, Democratic People's Republic of",
  "Laos" = "Lao People's Democratic Republic",
  "Macao" = "Macau",
  "North Macedonia" = "Macedonia",
  "Micronesia" = "Micronesia, Federated States of",
  "Moldova" = "Moldova, Republic of",
  "Caribbean Netherlands" = "Netherlands Antilles",
  "State of Palestine" = "Palestinian Territory, Occupied",
  "Réunion" = "Reunion",
  "Russia" = "Russian Federation",
  "Saint Kitts & Nevis" = "Saint Kitts and Nevis",
  "Saint Pierre & Miquelon" = "Saint Pierre and Miquelon",
  "St. Vincent & Grenadines" = "Saint Vincent and the Grenadines",
  "Sao Tome & Principe" = "Sao Tome and Principe",
  "Eswatini" = "Swaziland",
  "Tanzania" = "Tanzania, United Republic of",
  "Turks and Caicos" = "Turks and Caicos Islands",
  "United States" = "United States of America",
  "U.S. Virgin Islands" = "Virgin Islands, U.S.",
  "Wallis & Futuna" = "Wallis and Futuna"

)
Mapsui_Ready$Country <- dplyr::recode(Mapsui_Ready$Country,
'Guernsey' = "Channel Islands", "Jersey" = "Channel Islands"
)
Ready_joined_populiacija <- left_join(Mapsui_Ready, populiacija,
by = c('Country'='Salis'))
Ready_joined_populiacija <- Ready_joined_populiacija[-89,]
options(scipen = 999)
Ready_joined_populiacija <- Ready_joined_populiacija %>%
mutate(infringements_2 = Infringements/Populiacija*100000)
Ready_joined_populiacija$infringements_2 <-
round(Ready_joined_populiacija$infringements_2,2)
summary_Ready <- summary(Ready_joined_populiacija)
WorldData=map_data("world")
```

```
#pervadiname savo duomenis
Ready_joined_populiacija$Country <-dplyr::recode
(Ready_joined_populiacija$Country,
  'Antigua and Barbuda'='Barbuda','Brunei Darussalam'='Brunei',
  "Cote D'Ivoire" = 'Ivory Coast',
'Congo, The Democratic Republic of the' = 'Democratic Republic of the Congo',
'Congo' = 'Republic of Congo',
'Falkland Islands (Malvinas)' = 'Falkland Islands',
  'Micronesia, Federated States of' = 'Micronesia',
  'United Kingdom' = 'UK',
  "Channel Islands" = "Guernsey",
  'Iran, Islamic Republic of' = 'Iran',
  'Saint Kitts and Nevis' = 'Saint Kitts',
  'Korea, Republic of' = 'South Korea',
  "Lao People's Democratic Republic" = 'Laos',
  'Moldova, Republic of' = 'Moldova',
  "Netherlands Antilles" = "Bonaire",
  'Russian Federation' = 'Russia',
  'East Timor' = 'Timor-Leste',
  'Trinidad and Tobago' = 'Trinidad',
  'Tanzania, United Republic of' = 'Tanzania',
  'United States of America' = 'USA',
  'Holy See (Vatican City State)' = 'Vatican',
  'Saint Vincent and the Grenadines' = 'Grenadines',
  "Korea, Democratic People's Republic of" = "North Korea",
  'Palestinian Territory, Occupied' = 'Palestine',
  'Virgin Islands'= 'British Virgin Islands',
  'Virgin Islands'='Virgin Islands, U.S.'
)
#pervadiname R paketo duomenis
WorldData$region <- dplyr::recode(WorldData$region, 'Antigua' = 'Barbuda',
 'Nevis' = 'Saint Kitts',
 'Tobago' = 'Trinidad',
"Jersey" = "Guernsey"
)
Combined = WorldData[Ready_joined_populiacija$Country %in%
Ready_joined_populiacija$Country, ]
Combined$Infringements_2 =
Ready_joined_populiacija$infringements_2[
```

```
match(Combined$region, Ready_joined_populiacija$Country)]
number <- Ready_joined_populiacija %>%
subset(Country == "Hong Kong", select = infringements_2)
number_virgin_british <- Ready_joined_populiacija
%>% subset(Country == "British Virgin Islands", select = infringements_2)
number__virgin_us <- Ready_joined_populiacija %>%
subset(Country == "Virgin Islands, U.S.", select = infringements_2)
Combined <- Combined %>% mutate(Infringements_2 =
replace(Infringements_2, subregion == "Hong Kong", number))
virgin_number <- number_virgin_british+number__virgin_us
Combined <- Combined %>% mutate(Infringements_2 =
replace(Infringements_2, region == "Virgin Islands", virgin_number))
Combined$Infringements_2 <- as.numeric(Combined$Infringements_2)
Countries = unique(Combined$region)
CDF = data.frame(label1=Countries)
ggplot(Combined, aes(x=long, y=lat, group = group, fill = Infringements_2)) +
  geom_polygon(colour = "white") +
  scale_fill_continuous(low = "blue",
high = "red",
guide="colorbar") +
  theme_bw()  +
  labs(fill = "Pažeidimų skaičius 10000 gyventojų" ,
  title = "Šalių pažeidimų skaičius", x="", y="") +
  scale_y_continuous(breaks=c()) +
  scale_x_continuous(breaks=c()) +
  theme(panel.border =  element_blank())
#Šalinamos išskirtys
Combined <- Combined %>% mutate(Infringements_2 =
replace(Infringements_2, region == "Falkland Islands", 0))
Combined <- Combined %>% mutate(Infringements_2 =
replace(Infringements_2, region == "Antarctica", 0))
ggplot(Combined, aes(x=long, y=lat, group = group,
fill = Infringements_2)) +
  geom_polygon(colour = "white") +
  scale_fill_continuous(low = "blue",
high = "red",
guide="colorbar") +
  theme_bw()  +
  labs(fill = "Pažeidimų skaičius 100 tūkst. gyventojų" ,
```

```
  title = "Ready pažeidimų skaičius", x="", y="") +
  scale_y_continuous(breaks=c()) +
  scale_x_continuous(breaks=c()) +
  theme(panel.border =  element_blank())


#Dižiausios reikšmės
Ready_joined_populiacija = arrange(Ready_joined_populiacija,
desc(infringements_2), infringements_2)
head_Ready <- head(Ready_joined_populiacija,10)
head_Ready <- head_Ready[,c(1,4)]
colnames(head_Ready) <- c("Šalis", "Pažeidimų skaičius 10000 gyventojų")
kable(head_Ready, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down",full_width = F)


#Mažiausios reikšmės
Ready_joined_populiacija <- na.omit(Ready_joined_populiacija)
tail_Ready <- tail(Ready_joined_populiacija,10)
tail_Ready <- tail_Ready[,c(1,4)]
colnames(tail_Ready) <- c("Šalis", "Pažeidimų skaičius 10000 gyventojų")
kable(tail_Ready, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down",full_width = F)


#Sąryšio tarp pažeidimų internete ir BVP vienam gyventojui
kable(pajamu_grupes, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="striped")
Beast_joined_populiacija <- left_join(Mapsui_Beast, populiacija,
by = c('Country'='Salis'))
Beast_joined_populiacija <- Beast_joined_populiacija[-90,]
options(scipen = 999)
Beast_joined_populiacija <- Beast_joined_populiacija %>%
mutate(infringements_2 = Infringements/Populiacija*100000)
Beast_joined_populiacija$infringements_2 <-
round(Beast_joined_populiacija$infringements_2,2)
pajamu_lygis_salys$Economy <- dplyr::recode(pajamu_lygis_salys$Economy,
 'Bahamas, The'='Bahamas',
 'Cabo Verde'='Cape Verde',
```

```
 "Congo, Dem. Rep."="Congo, The Democratic Republic of the",
 "Congo, Rep."="Congo",
 "Côte d'Ivoire"="Cote D'Ivoire",
 "Curaçao"="Curacao",
 "Egypt, Arab Rep."="Egypt",
 "Eswatini" ="Swaziland",
 "Gambia, The" = "Gambia",
 "Hong Kong SAR, China" = "Hong Kong",
 "Iran, Islamic Rep." = "Iran, Islamic Republic of",
 "Korea, Dem. People's Rep."="Korea, Democratic People's Republic of",
 "Korea, Rep." = "Korea, Republic of",
 "Kyrgyz Republic" = "Kyrgyzstan",
 "Lao PDR" = "Lao People's Democratic Republic",
 "Macao SAR, China" = "Macau",
 "Micronesia, Fed. Sts." = "Micronesia, Federated States of",
 "Moldova" = "Moldova, Republic of",
 "São Tomé and Principe" = "Sao Tome and Principe",
 "Sint Maarten (Dutch part)" = "Sint Maarten",
 "Slovak Republic" = "Slovakia",
 "St. Kitts and Nevis" = "Saint Kitts and Nevis",
 "St. Lucia" = "Saint Lucia",
 "St. Martin (French part)" = "Saint Martin",
 "St. Vincent and the Grenadines" = "Saint Vincent and the Grenadines",
 "Syrian Arab Republic" = "Syria",
 "Taiwan, China" = "Taiwan",
 "Tanzania" = "Tanzania, United Republic of",
 "Timor-Leste" = "East Timor",
 "United States" = "United States of America",
 "Venezuela, RB" = "Venezuela",
 "Virgin Islands (U.S.)" = "Virgin Islands, U.S.",
 "Yemen, Rep." = "Yemen",
 "North Macedonia" = "Macedonia"

)
pajamu_grupes_join <- left_join(pajamu_lygis_salys,
Beast_joined_populiacija, by = c('Economy' = 'Country'))
pajamu_grupes_join <- na.omit(pajamu_grupes_join)
pajamu_grupes_join <- pajamu_grupes_join[,c(1,2,3,6)]
data_pajamos <- melt(pajamu_grupes_join)
```

```
#Pažeidimų charakteristikos - vidurkis ir variacija
Ready_joined_populiacija <-
left_join(Mapsui_Ready, populiacija, by = c('Country'='Salis'))
Ready_joined_populiacija <- Ready_joined_populiacija[-89,]
options(scipen = 999)
Ready_joined_populiacija <-
Ready_joined_populiacija %>%
mutate(infringements_2 = Infringements/Populiacija*100000)
Ready_joined_populiacija$infringements_2 <-
round(Ready_joined_populiacija$infringements_2,2)
summary_Ready <- summary(Ready_joined_populiacija)
Ready_joined_populiacija$Country <-
dplyr::recode(Ready_joined_populiacija$Country,
'Barbuda'= 'Antigua and Barbuda',
'Brunei'='Brunei Darussalam',
'Ivory Coast'="Cote D'Ivoire",
'Democratic Republic of the Congo'='Congo, The Democratic Republic of the',
'Republic of Congo'='Congo',
'Falkland Islands'= 'Falkland Islands (Malvinas)',
'Micronesia'='Micronesia, Federated States of',
'UK'='United Kingdom',
"Guernsey"="Channel Islands",
'Iran'='Iran, Islamic Republic of',
'Saint Kitts'='Saint Kitts and Nevis',
'South Korea'='Korea, Republic of',
'Laos'="Lao People's Democratic Republic",
'Moldova'='Moldova, Republic of',
"Bonaire"="Netherlands Antilles",
'Russia'='Russian Federation',
'Timor-Leste'='East Timor',
'Trinidad'='Trinidad and Tobago',
'Tanzania'='Tanzania, United Republic of',
'USA'='United States of America',
'Vatican'='Holy See (Vatican City State)',
'Grenadines'='Saint Vincent and the Grenadines',
"North Korea"= "Korea, Democratic People's Republic of",
'Palestine'='Palestinian Territory, Occupied',
```

```
)

data_pajamos_joined <- left_join(data_pajamos,
Ready_joined_populiacija, by = c('Economy' = 'Country'))
data_pajamos_joined$`Income group`<-
ordered(data_pajamos_joined$`Income group`,
  levels = c("Low income", "Lower middle income",
  "Upper middle income", "High income"))
colnames(data_pajamos_joined) <- c("Economy",
"Code","Income group","variable","Beast",
"Infringements","Populiacija", "Ready")
data_pajamos_joined <- data_pajamos_joined[,c(1,3,5,8)]
bendra <- group_by(data_pajamos_joined, `Income group`) %>%
  dplyr::summarise(
 count = n(),
 mean_Beast = mean(Beast, na.rm = TRUE),
 sd_Beast = sd(Beast, na.rm = TRUE),
 mean_Ready = mean(Ready, na.rm = TRUE),
 sd_Ready = sd(Ready, na.rm = TRUE),
  )
colnames(bendra) <- c("Pajamų grupė",
"Šalių skaičius", "Vidurkis Beast", "Standartinis nuokrypis Beast",
"Vidurkis Ready", "Standartinis nuokrypis Ready" )
bendra$`Pajamų grupė` <- c("Žemos pajamos","Žemesnės-vidutinės pajamos",
"Auskštesnės-vidutinės pajamos","Aukštos pajamos")
bendra$`Vidurkis Beast` <- round(bendra$`Vidurkis Beast`,2)
bendra$`Vidurkis Ready` <- round(bendra$`Vidurkis Ready`,2)
bendra$`Standartinis nuokrypis Beast` <-
round(bendra$`Standartinis nuokrypis Beast`,2)
bendra$`Standartinis nuokrypis Ready` <-
round(bendra$`Standartinis nuokrypis Ready`,2)
kable(bendra, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="striped")
procenta_pajamos <- bendra$`Šalių skaičius`/216*100


#ANOVA
colnames(data_pajamos_joined) <- c("Economy","group","Beast","Ready")
library(rstatix)
```

```
log_duomenys <- as.data.frame(cbind(data_pajamos_joined$Beast,
data_pajamos_joined$Ready))
log_duomenys <- log_duomenys %>% mutate(group = data_pajamos_joined$group)
log_duomenys <- log_duomenys %>%
mutate(Economy = data_pajamos_joined$Economy)
colnames(log_duomenys) <- c("Beast", "Ready", "group","Economy")
log_duomenys$Beast <- log(log_duomenys$Beast)
log_duomenys$Ready <- log(log_duomenys$Ready)
log_duomenys <-
  log_duomenys %>% mutate(bendri =
  data_pajamos_joined$Ready+data_pajamos_joined$Beast)
log_duomenys <-
  log_duomenys %>% mutate(log_bendri = log(bendri))
log_duomenys %>%
  group_by(group) %>%
  get_summary_stats(bendri, type = "mean_sd")
log_duomenys %>%
  group_by(group) %>%
  get_summary_stats(bendri, type = "mean_sd")

#Braižome grafikus:
pervadinti_log_duomenys <- log_duomenys
levels(pervadinti_log_duomenys$group) <-
c("Žemos", "Žemesnės-vidutinės","Aukštesnės vidutinės","Aukštos pajamos")
colnames(pervadinti_log_duomenys)[3] <- "Grupė"
ggpubr::ggboxplot(pervadinti_log_duomenys, x = "Grupė", y ="bendri",
 color = "Grupė", palette = c("#00AFBB", "#E7B800", "#FC4E07", "#FC4E07"),
 order = c("Žemos", "Žemesnės-vidutinės","Aukštesnės vidutinės",
 "Aukštos pajamos"),ylab = "Bendras pažeidimų sk.", xlab = "Pajamų grupė")
ggpubr::ggboxplot(pervadinti_log_duomenys, x = "Grupė", y ="log_bendri",
 color = "Grupė", palette = c("#00AFBB", "#E7B800", "#FC4E07", "#FC4E07"),
 order = c("Žemos", "Žemesnės-vidutinės","Aukštesnės vidutinės",
 "Aukštos pajamos"),ylab = "Bendras pažeidimų sk.", xlab = "Pajamų grupė")

#Išskirčių apžvalga duomenyse ir grafikai pašalinus didžiausias išskirtis
library(rstatix)
log_duomenys %>%
  group_by(group) %>%
  identify_outliers(bendri)
```

```
log_duomenys %>%
  group_by(group) %>%
  identify_outliers(log_bendri)
isskirtys_bendri <- log_duomenys %>%
  group_by(group) %>%
  identify_outliers(bendri)
isskirtys_bendri <- isskirtys_bendri[isskirtys_bendri[,7] == TRUE,]
isskirtys_bendri_log <- log_duomenys %>%
  group_by(group) %>%
  identify_outliers(log_bendri)
isskirtys_bendri_log <- isskirtys_bendri_log[isskirtys_bendri_log
[,7] == TRUE,]
data_pajamos_joined_bendri_outliers <- log_duomenys[-which
(log_duomenys$Economy %in% isskirtys_bendri$Economy),]
data_pajamos_joined_bendri_outliers_log <- log_duomenys[
-which(log_duomenys$Economy %in% isskirtys_bendri_log$Economy),]
ggpubr::ggboxplot(data_pajamos_joined_bendri_outliers_log, x = "group",
y ="log_bendri",
 color = "group", palette = c("#00AFBB", "#E7B800", "#FC4E07", "#FC4E07"),
 order = c("Low income", "Lower middle income", "Upper middle income",
 "High income"), ylab = "Pažeidimai log bendri", xlab = "Pajamų grupė")

#Normalumo tikrinimas
log_duomenys %>%
  group_by(group) %>%
  shapiro_test(log_bendri)
data_pajamos_joined_bendri_outliers_log %>%
  group_by(group) %>%
  shapiro_test(log_bendri)
ggpubr::ggqqplot(data_pajamos_joined_bendri_outliers, "log_bendri",
facet.by = "group")
ggpubr::ggqqplot(data_pajamos_joined_bendri_outliers_log, "log_bendri",
facet.by = "group")
ggpubr::ggqqplot(data_pajamos_joined_bendri_outliers, "bendri",
facet.by = "group")

#Homogeniškumas
data_pajamos_joined_bendri_outliers %>% levene_test(bendri ~ group)
log_duomenys %>% levene_test(bendri ~ group)
```

```
data_pajamos_joined_bendri_outliers_log %>% group_by("Grupė")%>%
levene_test(log_bendri ~ group)
pervadinti_log_duomenys %>% group_by("Grupė")%>%
levene_test(log_bendri ~ Grupė)

#tikrinama ANOVA
fit_anova_bendri <- aov(bendri ~ group, data=log_duomenys)
summary(fit_anova_bendri)
print(model.tables(fit_anova_bendri,"means"),digits=3)
#be iskirciu
fit_anova_bendri_log <- aov(log_bendri ~ group, data=log_duomenys)
summary(fit_anova_bendri_log)
print(model.tables(fit_anova_bendri_log,"means"),digits=3)
fit_anova_bendri_outliers <- aov(bendri ~ group, d
ata=data_pajamos_joined_bendri_outliers)
summary(fit_anova_bendri_outliers)
print(model.tables(fit_anova_bendri_outliers,"means"),digits=3)
fit_anova_bendri_outliers_log <- aov(log_bendri ~ group,
data=data_pajamos_joined_bendri_outliers_log)
summary(fit_anova_bendri_outliers_log)
print(model.tables(fit_anova_bendri_outliers_log,"means"),digits=3)
oneway.test(log_bendri ~ Grupė, data = pervadinti_log_duomenys)
oneway.test(log_bendri ~ group, data =
data_pajamos_joined_bendri_outliers_log)

#Games-Howell
library(userfriendlyscience)
one.way <- oneway(data_pajamos_joined_bendri_outliers_log$group,
y = data_pajamos_joined_bendri_outliers_log$log_bendri,
posthoc = 'games-howell')
one.way_2 <- oneway(pervadinti_log_duomenys$Grupė, y =
pervadinti_log_duomenys$log_bendri, posthoc = 'games-howell')

#Tikriname homogeniškumą likučiams (residuals)
plot(fit_anova_bendri_outliers_log, 1)
plot(fit_anova_bendri, 1)
plot(fit_anova_bendri_log, 1)

#Atliekame Levene testą tikrinti homogeniškumą
```

```
library(car)
leveneTest(bendri ~ group, data = log_duomenys)
leveneTest(log_bendri ~ group, data = log_duomenys)


#ANOVA testas be prielaidos apie variacijų lygybę ir liekanų testas.
oneway.test(bendri ~ group, data = log_duomenys)
plot(fit_anova_bendri, 2)
oneway.test(log_bendri ~ group, data = log_duomenys)
plot(fit_anova_bendri_log, 2)
aov_residuals_bendri <- residuals(object = fit_anova_bendri )
aov_residuals_bendri_log <- residuals(object = fit_anova_bendri_log )
oneway.test(bendri ~ group, data = data_pajamos_joined_bendri_outliers)
plot(fit_anova_bendri_outliers, 2)
oneway.test(log_bendri ~ group, data =
data_pajamos_joined_bendri_outliers_log)
plot(fit_anova_bendri_outliers_log, 2)
aov_residuals_bendri_outliers <-
residuals(object = fit_anova_bendri_outliers )
aov_residuals_bendri_outliers_log <- residuals(object =
fit_anova_bendri_outliers_log )

#Shapiro-Wilk testas
shapiro.test(x = aov_residuals_bendri_outliers )
kruskal.test(bendri ~ group, data = log_duomenys)
shapiro.test(x = aov_residuals_bendri_outliers_log )
kruskal.test(log_bendri ~ group, data = log_duomenys)
kruskal.test(log_bendri ~ group,
data = data_pajamos_joined_bendri_outliers_log)

#Post-Hoc po Kruskal Wallis reikšmingumo
#Dunn testas
library(FSA)

PT = dunnTest(log_bendri ~ group,
data=log_duomenys,
method="bh")
PT = PT$res
library(rcompanion)
```

```
cldList(comparison = PT$Comparison,
  p.value = PT$P.adj,
  threshold  = 0.05)
PT = dunnTest(log_bendri ~ group,
data=data_pajamos_joined_bendri_outliers_log,
method="bh")
PT = PT$res


#GNI duomenys gyventojui ir Ready/Beast filmų pažeidimų koreliacija
su pajamomis
library(car)
options(scipen = 999)
colnames(GNI) <- c("Country", "Code","Indicator", "gni")
GNI <- left_join(GNI, populiacija, by = c("Country" = "Salis"))
GNI <- na.omit(GNI)
#GNI vienam gyventojui
gni_per_capita =GNI$gni/GNI$Populiacija
GNI <- cbind(GNI,gni_per_capita)
data_pajamos_7 <- data_pajamos_joined
GNI <- left_join(GNI, data_pajamos_7, by = c("Country" = "Economy"))
GNI <- GNI[,c(-2,-3,-4)]
log_gni_per_capita <- log(GNI$gni_per_capita)
log_pazeidimai_Ready <- log(GNI$Ready)
log_pazeidimai_Beast <- log(GNI$Beast)
scatterplot(log_gni_per_capita,GNI$Ready)
scatterplot(log_gni_per_capita,GNI$Beast)
plot(density(GNI$Ready), main="Density Plot: pažeidimai Ready",
ylab="Frequency", sub=paste("Skewness:",
round(e1071::skewness(data_pajamos_joined$Ready), 2)))
plot(density(GNI$Beast), main="Density Plot: pažeidimai Ready",
ylab="Frequency", sub=paste("Skewness:", round(e1071::
skewness(data_pajamos_joined$Beast), 2)))
plot(density(log_gni_per_capita),
main="Density Plot: GNI per gyventoją",
ylab="Frequency", sub=paste("Skewness:", round(e1071::skewness
(data_pajamos_joined$Beast), 2)))
cor(log_gni_per_capita,GNI$Ready)
cor(log_gni_per_capita,GNI$Beast)
```

```
head(sort(GNI$gni_per_capita,decreasing=TRUE), n = 10)
head(arrange(GNI,(gni_per_capita)), n = 50)


#Nagrinėjamas pasiskirstymas pagal lytį šalyse ir BVP vienam gyventojui (USD)
GDP_euromonitor$Geography <- dplyr::recode(GDP_euromonitor$Geography,
'Cabo Verde'='Cape Verde',
'Congo, Democratic Republic'='Congo, The Democratic Republic of the',
'Congo-Brazzaville'="Congo",
"Côte d'Ivoire" = "Cote D'Ivoire",
"Eswatini" = "Swaziland",
"Hong Kong, China" = "Hong Kong",
"Iran" = "Iran, Islamic Republic of",
"North Korea" = "Korea, Democratic People's Republic of",
"South Korea" = "Korea, Republic of",
"Laos" = "Lao People's Democratic Republic",
"Macau, China"="Macau",
"Moldova" = "Moldova, Republic of",
"Russia" = "Russian Federation",
"Sao Tomé e Príncipe" = "Sao Tome and Principe",
"St Kitts and Nevis" = "Saint Kitts and Nevis",
"St Lucia" = "Saint Lucia",
"St Vincent and the Grenadines" = "Saint Vincent and the Grenadines",
"Tanzania" = "Tanzania, United Republic of",
"USA" = "United States of America",
"US Virgin Islands" = "Virgin Islands, U.S.",
"North Macedonia" = "Macedonia",
"Réunion" = "Reunion"

)
GDP_euromonitor <- left_join(GDP_euromonitor, populiacija,
by = c("Geography" = "Salis"))
GDP_euromonitor <- na.omit(GDP_euromonitor)
GDP_euromonitor <- GDP_euromonitor %>% mutate(BVP_vienam_gyventojui =
GDP_euromonitor$`2019`/GDP_euromonitor$Populiacija*1000000)
data_pajamos_2 <- left_join(data_pajamos_joined,
GDP_euromonitor, by = c('Economy' = 'Geography'))
data_pajamos_2 <- data_pajamos_2[,c(-5)]
grouped_2 <- data_pajamos_2 %>% group_by(Economy) %>%
```

```
tally(BVP_vienam_gyventojui)
colnames(grouped_2) <- c("Šalis","BVP vienam gyenojui")
tail_rouped_2 <- tail(grouped_2,10)
kable(tail_rouped_2, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="striped")
head_rouped_2 <- head(grouped_2,10)
kable(head_rouped_2, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="striped")


#Koreliacijas tarp Ready/Beast filmų pažeidimų skaičiaus 100 tūkst.
gyventojų sąryšis su BVP vienam gyventojui
data_pajamos_5 <- left_join(data_pajamos_4,
Lyties_pasisikrstymas_euromonitor,by = c('Economy' = 'Geography'))
lyties_procentai <- data_pajamos_5 %>% mutate(vyrai_proc =
(data_pajamos_5$Male*1000)/data_pajamos_5$Populiacija)
lyties_procentai <-lyties_procentai %>% mutate(moterys_proc =
(lyties_procentai$Female*1000)/data_pajamos_5$Populiacija)
cor.test(lyties_procentai$Ready,lyties_procentai$BVP_vienam_gyventojui,
method = "spearman")
cor.test(lyties_procentai$Ready,lyties_procentai$vyrai_proc,
method = "spearman")
cor.test(lyties_procentai$Beast,lyties_procentai$BVP_vienam_gyventojui,
method ="spearman")
cor.test(lyties_procentai$Beast,lyties_procentai$vyrai_proc,
method ="spearman")


#Apžvelgsime amžiaus grupių pasiskyrstymo šalyse duomenis iš
Euromonitor duomenų bazės
library(aod)
library(ggplot2)
library(reshape2)
melted <- lyties_procentai[,c(1,7:14)]
melted <- na.omit(melted)
melted_joined <- left_join(melted, populiacija, by = c('Economy' = 'Salis'))
lyties_procentai[,7:14] <- lyties_procentai[,7:14]/
lyties_procentai$Populiacija*100
pagrindine_lentele <- lyties_procentai
```

```
melted_joined[,2:9] <- melted_joined[,2:9]/melted_joined$Populiacija*100
summary_melted <- summary(melted_joined)
kable(summary_melted, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="striped")


#KLASTERINĖ ANALIZĖ:
#Pirmiausia apžvelgsime Ready ir Beast filmo duomenis
pagrindine_2 <- na.omit(pagrindine_lentele)
klasterine_1 <- pagrindine_2
klasterine_Beast <- klasterine_1[klasterine_1$Beast > 1095,]
klasterine_Ready <- klasterine_1[klasterine_1$Ready > 461,]
nr_Beast <- nrow(klasterine_Beast)
nr_Ready <- nrow(klasterine_Ready)
klasterine_Beast <- klasterine_Beast %>% mutate(index = 1:nr_Beast)
klasterine_Ready <- klasterine_Ready %>% mutate(index = 1:nr_Ready)
klasterine_Beast <- klasterine_Beast[moveme(names(klasterine_Beast),
"index first")]
klasterine_Ready <- klasterine_Ready[moveme(names(klasterine_Ready),
"index first")]
unique_salys_Beast <- unique(klasterine_Beast$Economy)
unique_salys_Ready <- unique(klasterine_Ready$Economy)
salos_Beast <- c("Antigua and Barbuda","Aruba","Bahamas",
"Barbados",
"Cape Verde", "Comoros","Curacao","Fiji","French Polynesia","Grenada",
"Guam","Kiribati","Madagascar","Maldives","Mauritius",
"New Caledonia","Samoa","Sao Tome and Principe",
"Seychelles","Solomon Islands","Sri Lanka","Saint Lucia",
"Saint Vincent and the Grenadines","Tonga","Vanuatu",
"Virgin Islands, U.S.","Jamaica","Puerto Rico",
"Trinidad and Tobago","Costa Rica", "Hong Kong",
"Brunei Darussalam", "Korea, Republic of")
salos_Ready <- c("Antigua and Barbuda","Aruba",
"Bahamas","Barbados",
"Cape Verde","Curacao","Fiji","French Polynesia","Grenada","Guam",
"Kiribati","Maldives","Mauritius","New Caledonia",
"Samoa","Sao Tome and Principe", "Seychelles",
"Sri Lanka","Saint Lucia","Saint Vincent and the Grenadines",
"Tonga","Vanuatu","Virgin Islands, U.S.",
```

```
"Jamaica","Puerto Rico","Trinidad and Tobago","Costa Rica",
"Brunei Darussalam",
"Korea, Republic of","Hong Kong","Macau")
klasterine_Beast <- Beast_filmas_daily[,c("Date",unique_salys_Beast)]
klasterine_Ready <- Ready_filmas_daily[,c("Date",unique_salys_Ready)]
klasterine_Beast_salos <- klasterine_Beast
[ , !(names(klasterine_Beast) %in% salos_Beast)]
klasterine_Ready_salos <- klasterine_Ready
[ , !(names(klasterine_Ready) %in% salos_Ready)]
melt_klasterine_Beast <- melt(klasterine_Beast,"Date")
melt_klasterine_Beast <- left_join(melt_klasterine_Beast,
populiacija,by =c("variable" = "Salis"))
melt_klasterine_Beast_salos <- melt(klasterine_Beast_salos,"Date")
melt_klasterine_Beast_salos <- left_join(melt_klasterine_Beast_salos,
populiacija,
by =c("variable" = "Salis"))
melt_klasterine_Ready <- melt(klasterine_Ready,"Date")
melt_klasterine_Ready <- left_join(melt_klasterine_Ready,
populiacija,by =c("variable" = "Salis"))
melt_klasterine_Ready_salos <- melt(klasterine_Ready_salos,"Date")
melt_klasterine_Ready_salos <- left_join(melt_klasterine_Ready_salos,
populiacija,
by =c("variable" = "Salis"))



#Žalios laiko eilutės Beast
library(gghighlight)
ggplot(melt_klasterine_Beast, aes(Date, value, color=variable)) +
  geom_line(stat="identity") +
  ylab("pažeidimai 100k žmonių") +
  gghighlight(max(value) > 160,
max_highlight = 4,
use_direct_label = TRUE) +
  theme_minimal() +
  theme(legend.position = 'none')

#Išbrėžiame grafiką
ggplot(melt_klasterine_Beast_salos, aes(Date, value, color=variable)) +
```

```
  geom_line(stat="identity") +
  ylab("Beast pažeidimai 100 tūkst. žmonių") +
  xlab("Data")+
  gghighlight(max(value) > 55,
max_highlight = 4,
use_direct_label = TRUE) +
  theme_minimal() +
  theme(legend.position = 'none')
infringement_per_100k_Beast_salos <- melt_klasterine_Beast_salos %>%
  dplyr::select(value, Date, variable) %>%
  drop_na()

#Atstumai
spread_h_per_100k_Beast_salos <- infringement_per_100k_Beast_salos %>%
  spread("variable", value)
infringements_Beast_salos <- t(spread_h_per_100k_Beast_salos[-1])
infringements_dist_Beast_salos <- dist(infringements_Beast_salos,
method="canberra")
fit_Beast_complete_salos <- hclust(infringements_dist_Beast_salos,
method="complete")
plot(fit_Beast_complete_salos, cex = 0.5, hang = -1, main = "Dendograma")
#Pasirenkame tinkamiausią metodą skaičiuoti tarp klasterių:
a <- c( "average", "single", "complete", "ward")
names(a) <- c( "average", "single", "complete", "ward")
ac <- function(x) {
  cluster::agnes(infringements_Beast_salos, method = x)$ac
}
map_dbl(a, ac)

#klasterių metodo parinkimas
library(factoextra)
factoextra::fviz_nbclust(infringements_Beast_salos,kmeans, method =
"wss")
factoextra::fviz_nbclust(infringements_Beast_salos, kmeans, method =
"silhouette")

#Dendograma
fit_Beast_ward_hclust_salos <- hclust(infringements_dist_Beast_salos,
method="ward")
```

```
plot(fit_Beast_ward_hclust_salos, cex = 0.6, hang = -1)
rect.hclust(fit_Beast_ward_hclust_salos, k=8, border="cadetblue")
clustered_data_Beast_salos <- cutree(fit_Beast_ward_hclust_salos, k=8)
clustered_data_tidy_Beast_salos <- as.data.frame(
as.table(clustered_data_Beast_salos)) %>% glimpse()
colnames(clustered_data_tidy_Beast_salos) <- c("variable","cluster")
clustered_data_tidy_Beast_salos$Country <-
as.character(clustered_data_tidy_Beast_salos$variable)
joined_clusters_Beast_salos <- melt_klasterine_Beast_salos %>%
  inner_join(clustered_data_tidy_Beast_salos, by = "variable") %>%
  glimpse()
table(clustered_data_tidy_Beast_salos$cluster)

#Vaizduojam klasterį
cluster1_Beast_salos <- joined_clusters_Beast_salos %>%
filter(cluster == "1")
ggplot(cluster1_Beast_salos, aes(Date, value)) +
  geom_line(color="grey") +
  theme_minimal() +
  ylab("Pažeidimai 100K žmonių") + xlab("") +
  geom_smooth(method="auto",color="red", se=F, size=0.5) +
  facet_wrap(~variable)

#LOGISTINĖ REGRESIJA
library(dplyr)
median_Beast <- median(pagrindine_2$Beast)
median_Ready <- median(pagrindine_2$Ready)
Logistine = pagrindine_2 %>%
  mutate(pazeidimai_Beast=ifelse(Beast >= median_Beast ,1,0))
Logistine$pazeidimai_Beast<-as.factor(Logistine$pazeidimai_Beast)
summary(Logistine$pazeidimai_Beast)
Logistine = Logistine %>%
  mutate(pazeidimai_Ready=ifelse(Ready >= median_Ready ,1,0))
Logistine$pazeidimai_Ready<-as.factor(Logistine$pazeidimai_Ready)
summary(Logistine$pazeidimai_Ready)
Logistine$group <- as.factor(Logistine$group)

#Konstruojam logistinę regresiją:
Logistine_2 <- Logistine
```

```
#logaritmuojam BVP
Logistine_2$BVP_vienam_gyventojui <- log(Logistine_2$BVP_vienam_gyventojui+1/2)
xtabs(~pazeidimai_Beast + group, data = Logistine_2)
logit_Beast <- glm(pazeidimai_Beast ~
BVP_vienam_gyventojui+ ‘30-34‘+ ‘35-39‘ +  ‘40-44‘  + ‘45-49‘+‘65+‘+
vyrai_proc+group  , data = Logistine_2, family = "binomial")
summary(logit_Beast)

#Antroji logistinė regresija
Regionas <- read_excel("C:/Users/Roberta/Desktop/Bakalauras/Duomenys_1/
Regionas.xlsx")
Logistine_2_regionas <- left_join(Logistine_2, Regionas, by =
c(’Economy’ = ’Country’))
Logistine_2_regionas <- unique(Logistine_2_regionas)
Logistine_2_regionas$‘World Region‘ <- as.factor(Logistine_2_regionas$
‘World Region‘)
xtabs(~pazeidimai_Beast + ‘World Region‘, data = Logistine_2_regionas)
logit_Beast_regionas <- glm(pazeidimai_Beast ~ BVP_vienam_gyventojui +
‘40-44‘  + vyrai_proc + ‘World Region‘, data = Logistine_2_regionas,
family = "binomial")
summary(logit_Beast_regionas)

#Susiskirstom duomenis į 2 grupes, tikrinam patikimumą prognozavimo
split <- caTools::sample.split(Logistine_2_regionas, SplitRatio = 0.8)
split
train <- subset(Logistine_2_regionas, split == "TRUE")
test <- subset(Logistine_2_regionas, split == "FALSE")
model <- glm(pazeidimai_Beast ~ BVP_vienam_gyventojui + ‘40-44‘  +
vyrai_proc + ‘World Region‘, data = train, family = "binomial")
res <- predict(model, test.type ="response")
res <- predict(model, train.type ="response")
confmatrix <- table(Actualvalue=train$pazeidimai_Beast,
Predictvalue = res > 0.5)
confmatrix
(confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)

#Tikrinam abu modelius ir lyginam R^2
ll.null <-logit_Beast$null.deviance/-2
ll.proposed <- logit_Beast$deviance/-2
```

```
(ll.null-ll.proposed)/ll.null
pchisq(2*(ll.null-ll.proposed),df=(length(logit_Beast$coefficients)-1))
ll.null <-logit_Beast_regionas$null.deviance/-2
ll.proposed <- logit_Beast_regionas$deviance/-2
(ll.null-ll.proposed)/ll.null
pchisq(2*(ll.null-ll.proposed),df=(
length(logit_Beast_regionas$coefficients)-1))
probabilities <- predict(logit_Beast_regionas, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, "1", "0")
head(predicted.classes)
mydata <- Logistine_2_regionas %>%
  dplyr::select_if(is.numeric)
predictors <- colnames(mydata)
mydata <- mydata %>%
  mutate(logit = log(probabilities/(1-probabilities))) %>%
  gather(key = "predictors", value = "predictor.value", -logit)
ggplot(mydata, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")
library(aod)
wald.test(b = coef(logit_Beast_regionas), Sigma =
vcov(logit_Beast_regionas), Terms = 4:6)
car::residualPlot(logit_Beast)
car::residualPlot(logit_Beast_regionas)

# Reikšmės, kurios turi įtaką (išskirtys)
plot(logit_Beast_regionas, which = 4, id.n = 3)
model.data <- augment(logit_Beast_regionas) %>%
  mutate(index = 1:n())
model.data %>% top_n(3, .cooksd)
ggplot(model.data, aes(index, .std.resid)) +
  geom_point(aes(color = pazeidimai_Beast), alpha = .5) +
  theme_bw()
model.data %>%
  filter(abs(.std.resid) > 3)

# Sezoniškumas
```

```
library(lubridate)
ts_Beast <- ts(klasterine_Beast$India,
 freq=365.25/7,
 start=decimal_date(ymd("2019-07-01")))
ts_Beast <- ts_Beast/145934462 *100000
par(mfrow = c(1,2))
plot.ts(ts_Beast, type ="l")
par(mfrow = c(1,2))
acf(ts_Beast)
pacf(ts_Beast)
dec_Beast <- decompose(ts_Beast)
plot(dec_Beast)
lm_Beast <- lm(ts_Beast ~ dec_Beast$seasonal + dec_Beast$trend)
summary(lm_Beast)
```