

TECHNICKÁ UNIVERZITA V KOŠICIACH

FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Objavovanie Znalostí

Zadanie

Ciele zadania

- Pochopenie dát
- Exploračná analýza dát
- Príprava dát
- Riešenie klasifikačnej/regresnej úlohy
- Riešenie popisnej úlohy

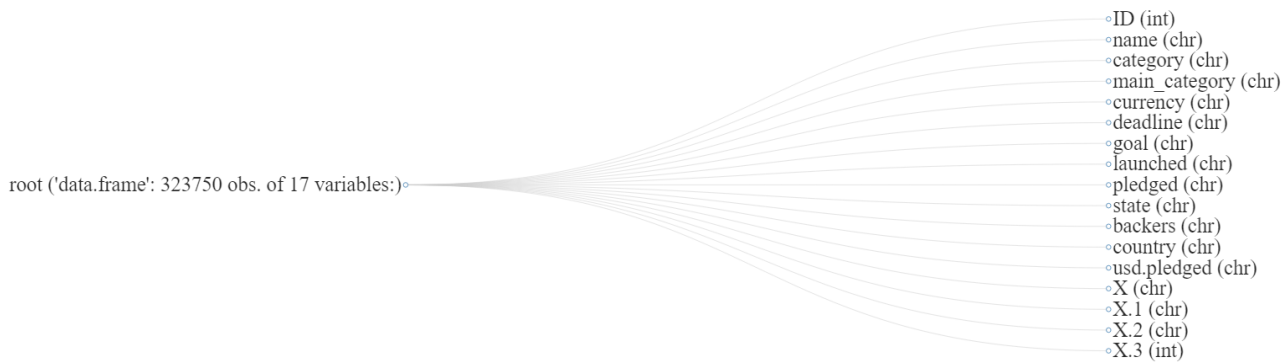
Dáta

Dáta zbrané z platformy Kickstarter.

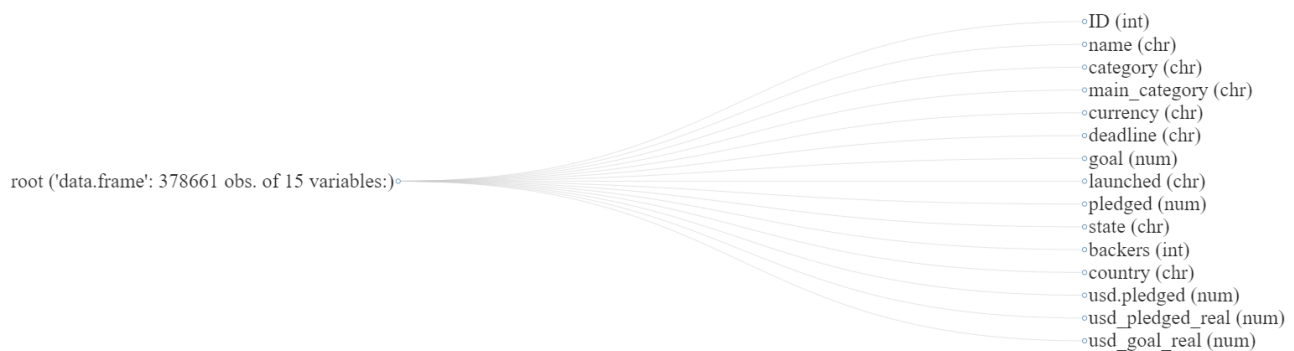
Kickstarter je stránka na získavanie financií na realizáciu kreatívnych, vedeckých a priemyselných projektov v rámci crowdfundingovej schémy (čiže dobrovoľné investovania). Kickstarter financuje rôzne projekty, ako napríklad tvorbu amerických nezávislých filmov, komiksov, videohier, hudby a ďalších.

Tabuľky:

- Kickstarter2016 - dáta z platformy kickstarter z roku 2016

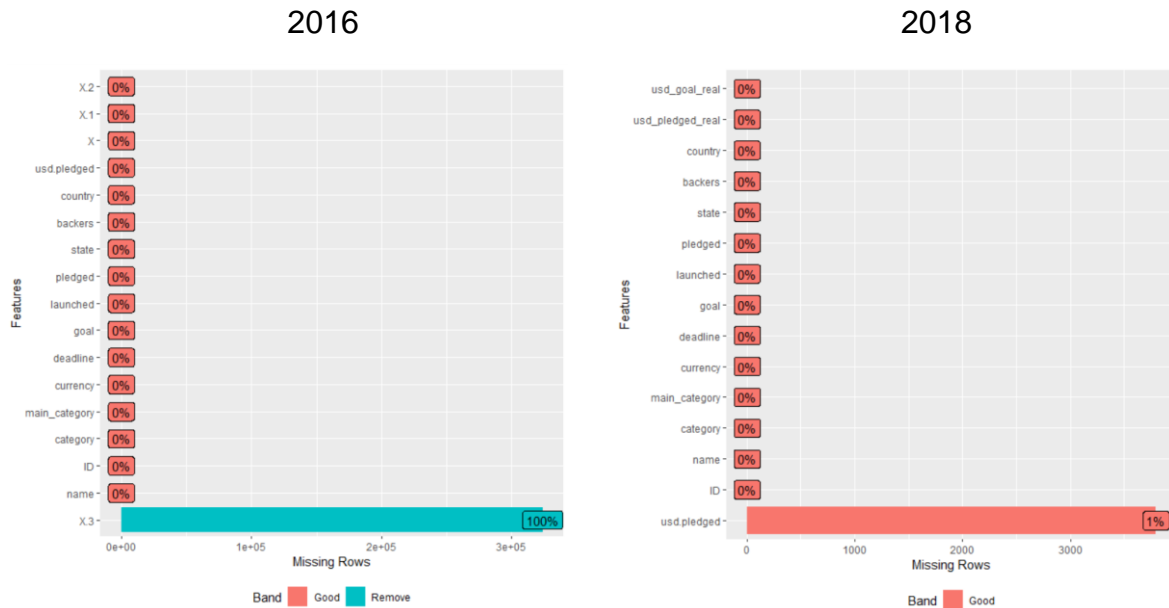


- Kickstarter2018 - dáta z platformy kickstarter z roku 2018



Pochopenie dát a Exploračná analýza dát

Úplnosť a kvalita dát:



Tabuľka Kickstarter2016 obsahovala viac chýbajúcich dát ako Kickstarter2018.

V tabuľke Kickstarter2018 obsahuje hodnotu NaN len stĺpec usd.pledged. Chýbajúce hodnoty by sme vedeli doplniť pomocou stĺpcov pledged a currency.

Tabuľka Kickstarter2016 obsahuje tiež jeden stĺpec obsahujúci NaN hodnoty. Keďže stĺpec obsahuje len NaN hodnoty je možné ho odstrániť.

Niektoré riadky obsahovali posunuté dáta a preto ich bolo nutné odstrániť.

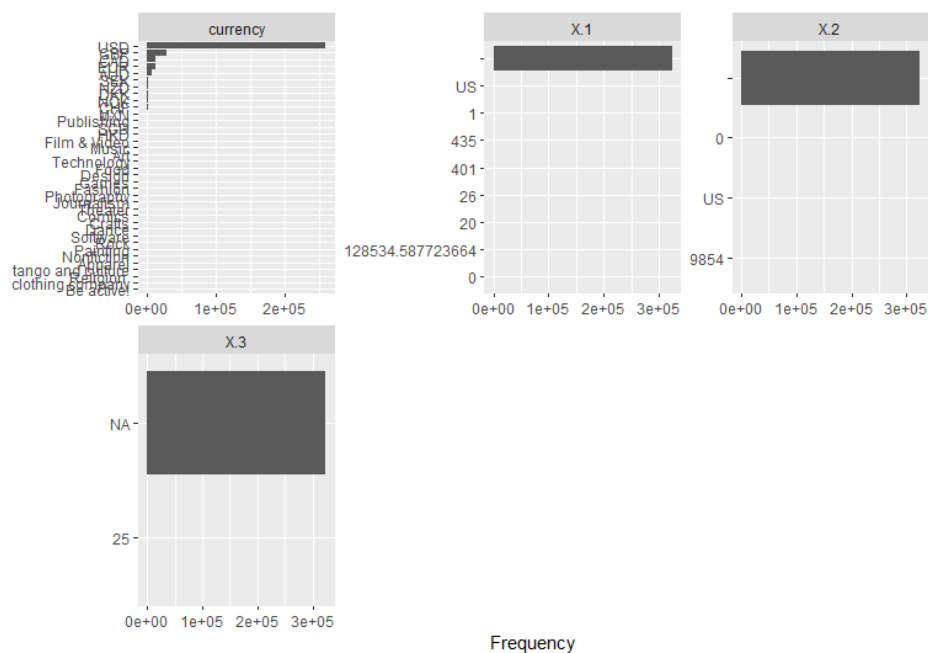
Využívané meny:

Oba datasety obsahovali rovnaké meny až na Japonský jen ktorý obsahoval len dataset Kickstarter2018.

- "USD" - Americký dolár
- "CAD" - Kanadský dolár
- "GBP" - Britská libra
- "AUD" - Austrálsky dolár
- "EUR" - Euro
- "SEK" - Švédská koruna
- "NZD" - Novozélandský dolár
- "DKK" - Dánska koruna
- "CHF" - Švajčiarsky frank
- "HKD" - Hongkonský dolár
- "SGD" - Singapurský dolár
- "NOK" - Nórska koruna
- "MXN" - Mexické peso
- "JPY" - Japonský jen
-

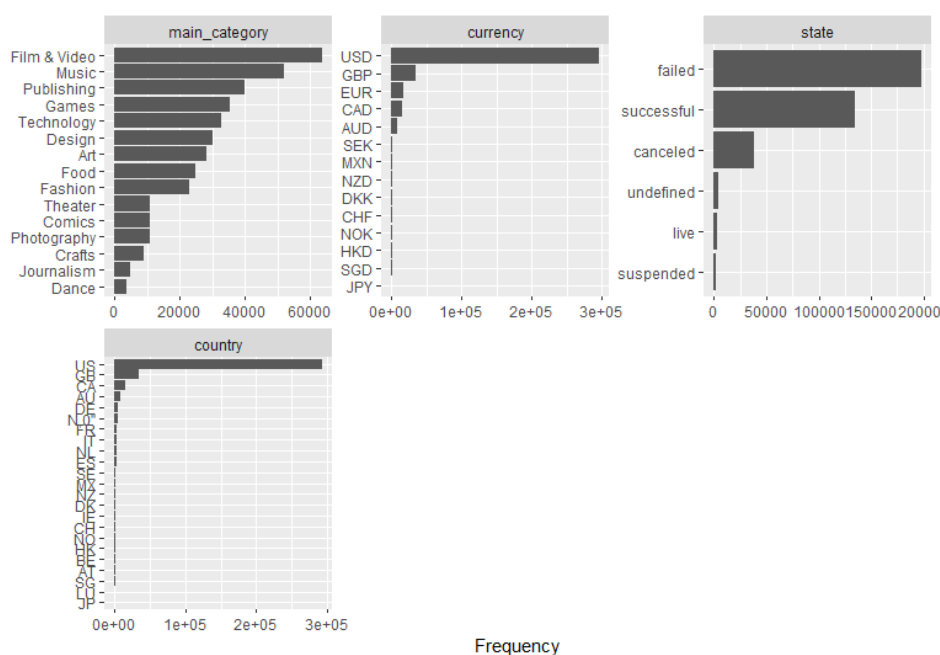
Pre pochopenie dát sme si vytvorili niekoľko barplotov.

Kickstarter2016



Z barplotu pre dáta z roku 2016 vidíme že v tomto datasete výrazne prevláda mena USD. Tiež si môžeme všimnúť, že v stĺpci “currency” je dátový posun, ktorý treba odstrániť a zistili sme, že sú stĺpce, obsahujúce len chýbajúce hodnoty.

Kickstarter2018



Z barplotu pre dáta z roku 2018 vidíme že i v tomto datasete výrazne prevláda mena USD. Pomocou tejto vizualizácie dát sme zistili že dataset obsahuje aj stavy ktoré by nám mohli zhoršiť tréning a preto ich možno odstrániť.

Pochopenie predspracovaných dát pred spojením tabuliek:

Kickstarter2016:

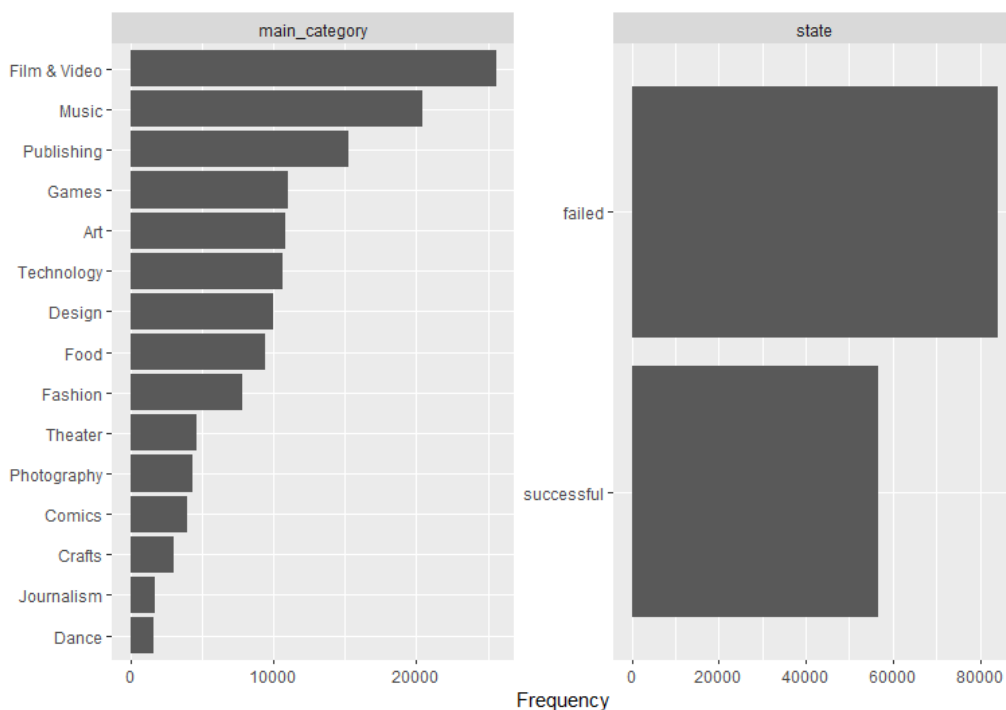
Korelácia medzi počtom podporovateľov a stavom

	backers	state_failed	state_successful
backers	1	-0.1153083	0.1153083
state_failed	-0.1153083	1	-1
state_successful	0.1153083	-1	1

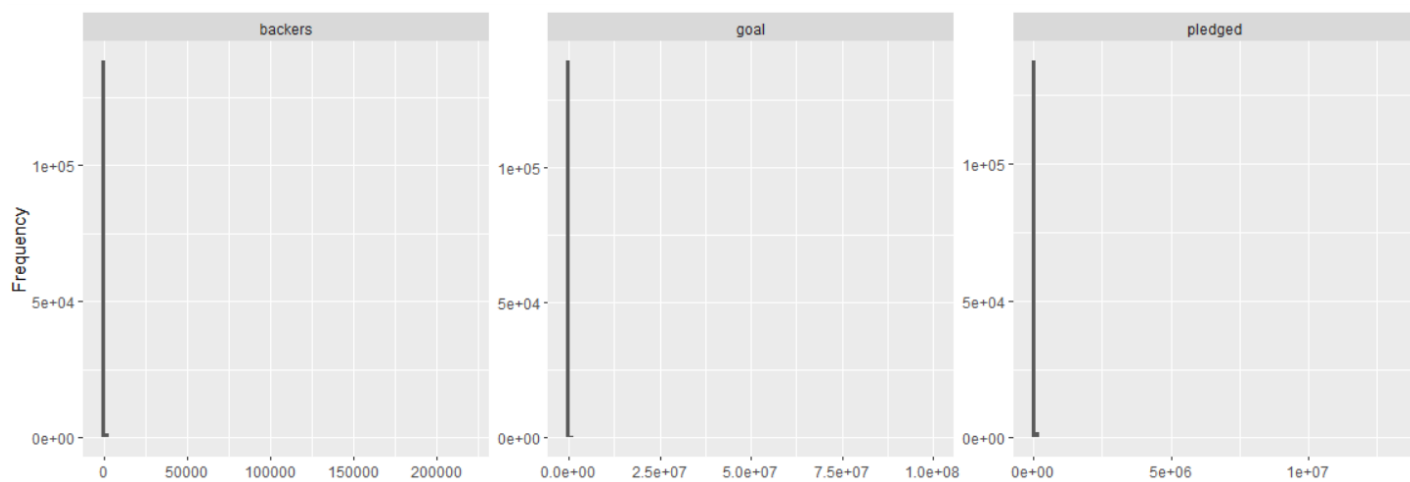
Sumarizácia dát z niektorých atribútov

goal	pledged	backers	state %
Min. 0	Min. 0	Min. 0	state_failed 0.5979
Prvý kvartil 2000	Prvý kvartil 50	Prvý kvartil 2	
Mean 39280	Mean 9161	Mean 111	state_successful 0.4021
Tretí kvartil 15000	Tretí kvartil 4429	Tretí kvartil 61	
Max. 100000000	Max. 13285226	Max 219382	

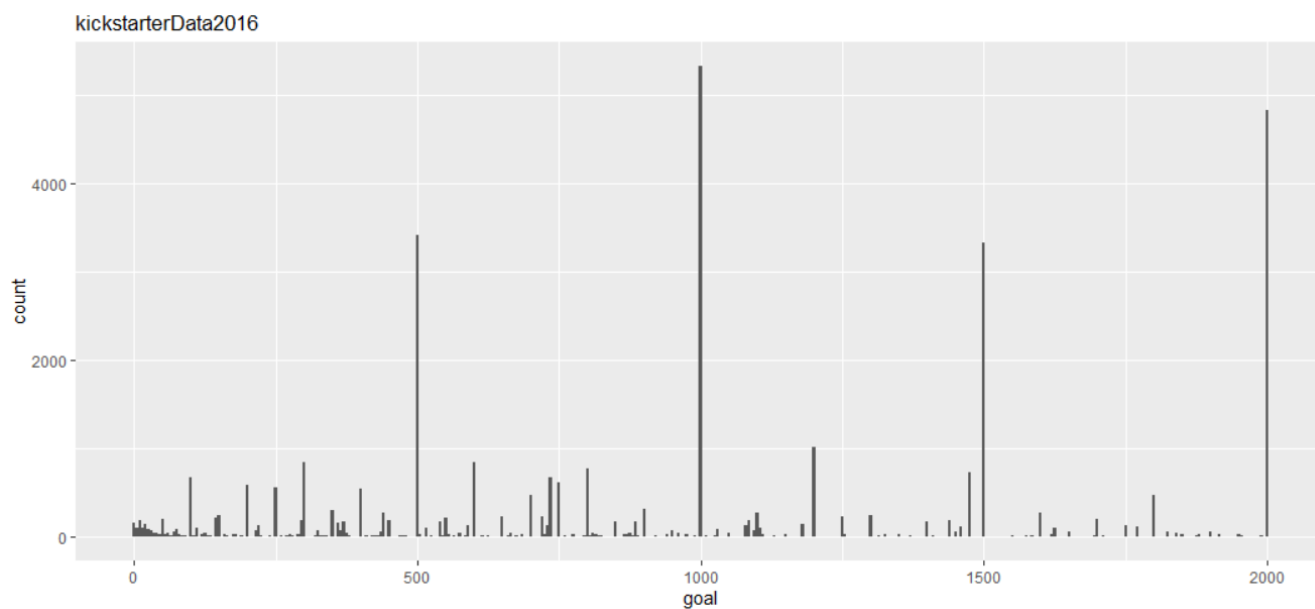
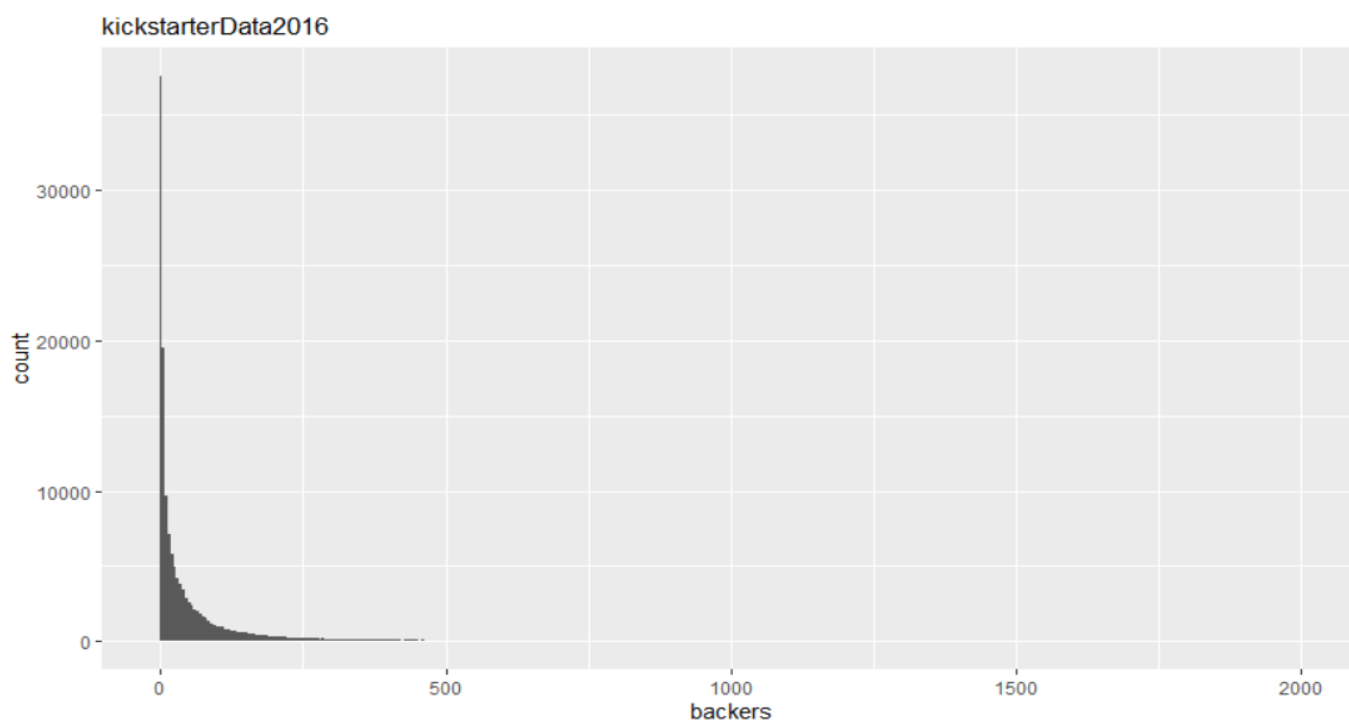
Zástupnosť tried v atribútoch “main_category” a “state”. Podľa dole uvedeného barplotu vidíme, že prevažuje kategória “Film&Video” a stav “failed”.

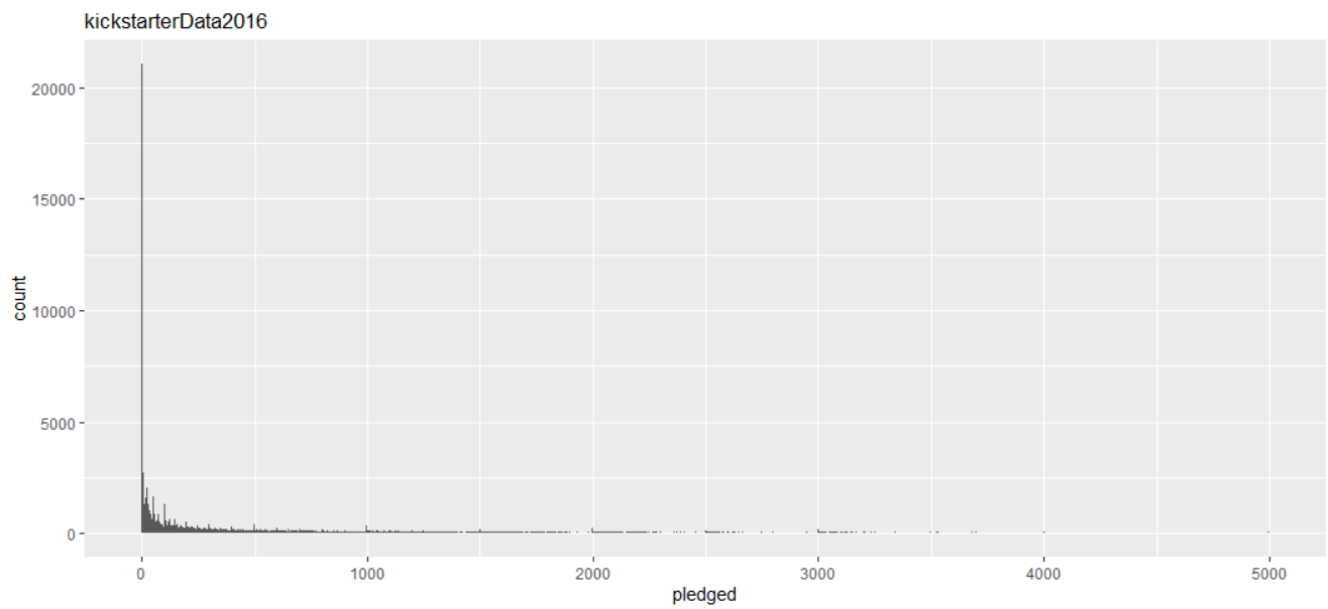


Distribúcia numerických atribútov. Podľa tejto distribúcie vidíme, že existujú nejaké špecifické prípady.

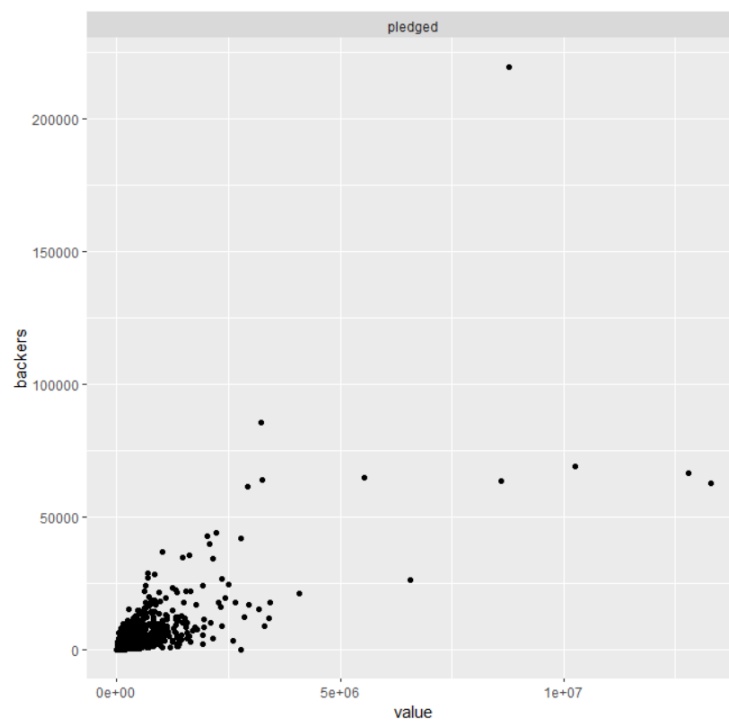


Vymazaním tých špecifických prípadov vidíme ako sa mení distribúcia.

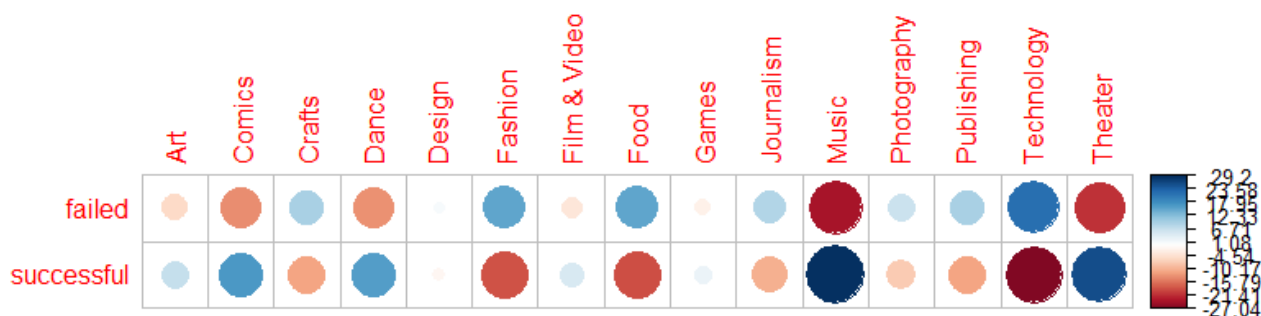




Rozptyl dát. Investované peniaze podľa počtu podporovateľov.



Chi-kvadrat test nezávislosti atribútov



Kickstarter2018:

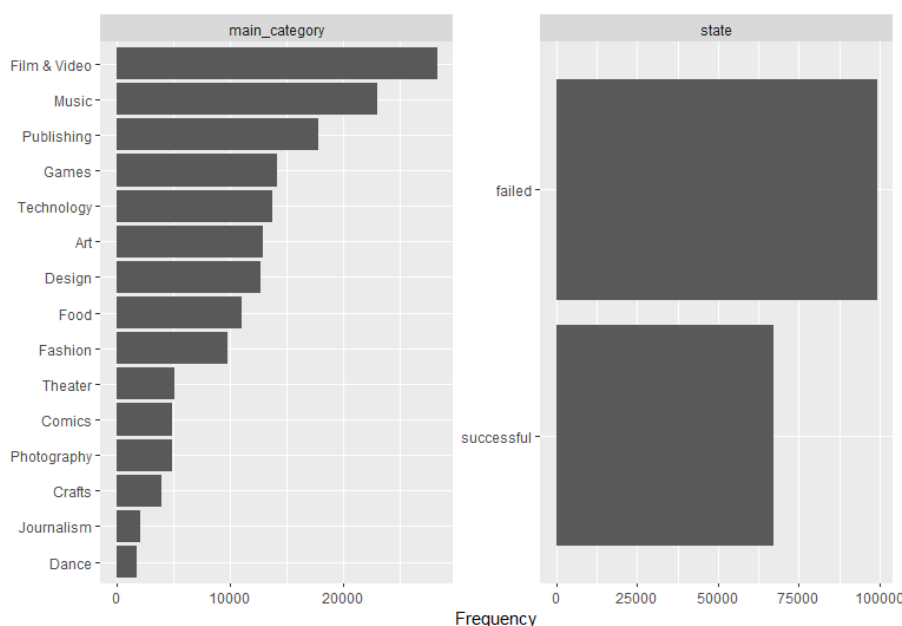
Korelácia medzi počtom podporovateľov a stavom

	backers	state_failed	state_successful
backers	1	-0.1155284	0.1155284
state_failed	-0.1155284	1	-1
state_successful	0.1155284	-1	1

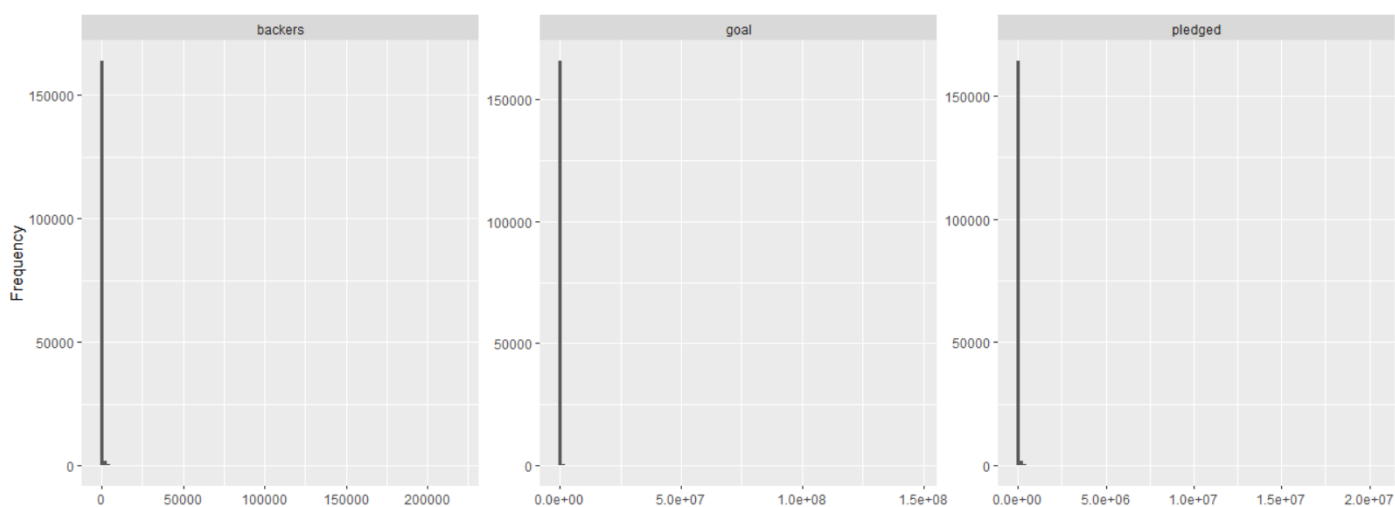
Sumarizácia dát z niektorých atribútov

goal	pledged	backers	state %
Min. 0	Min. 0	Min. 0	state_failed 0.5957
Prvý kvartil 2000	Prvý kvartil 50	Prvý kvartil 2	
Mean 42689	Mean 9161	Mean 119	state_successful 0.4043
Tretí kvartil 15000	Tretí kvartil 4613	Tretí kvartil 63	
Max. 147400000	Max. 20338986	Max 219382	

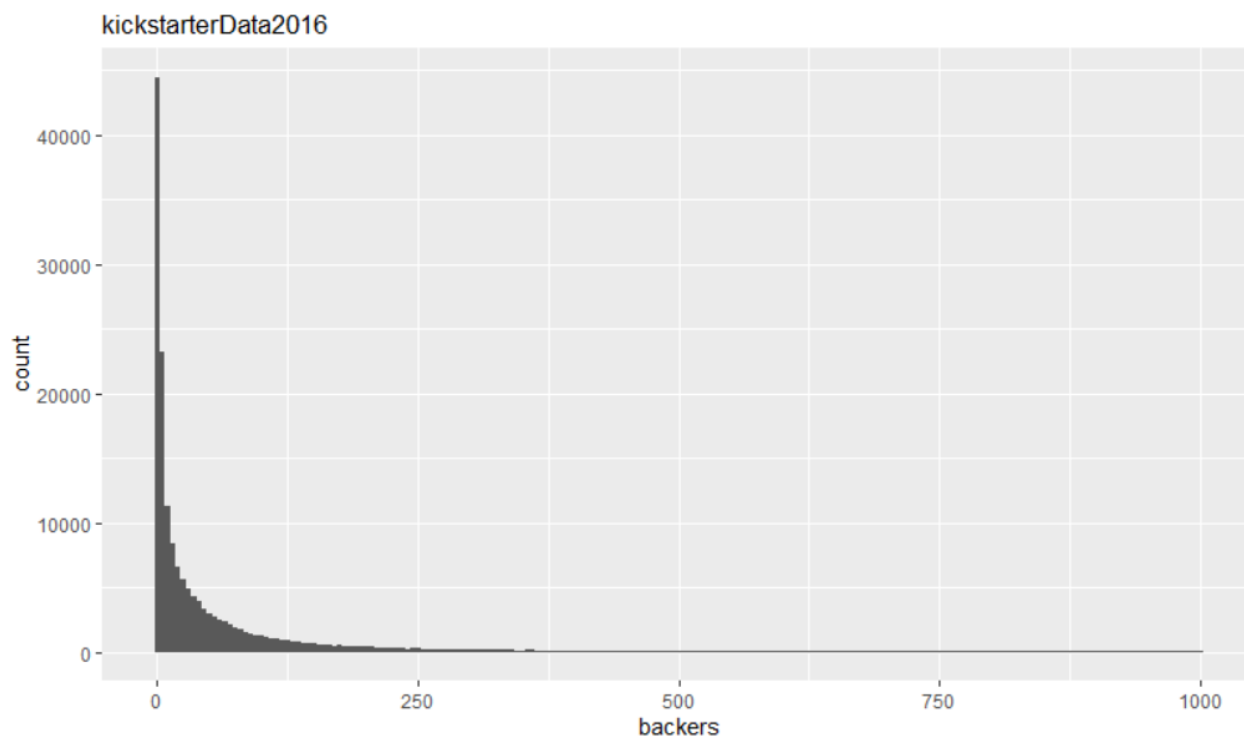
Zástupnosť tried v atribútoch “main_category” a “state”. Podľa dole uvedeného barplotu vidíme, že tak isto ako v roku 2016 prevažuje kategória “Film&Video” a stav “failed”.

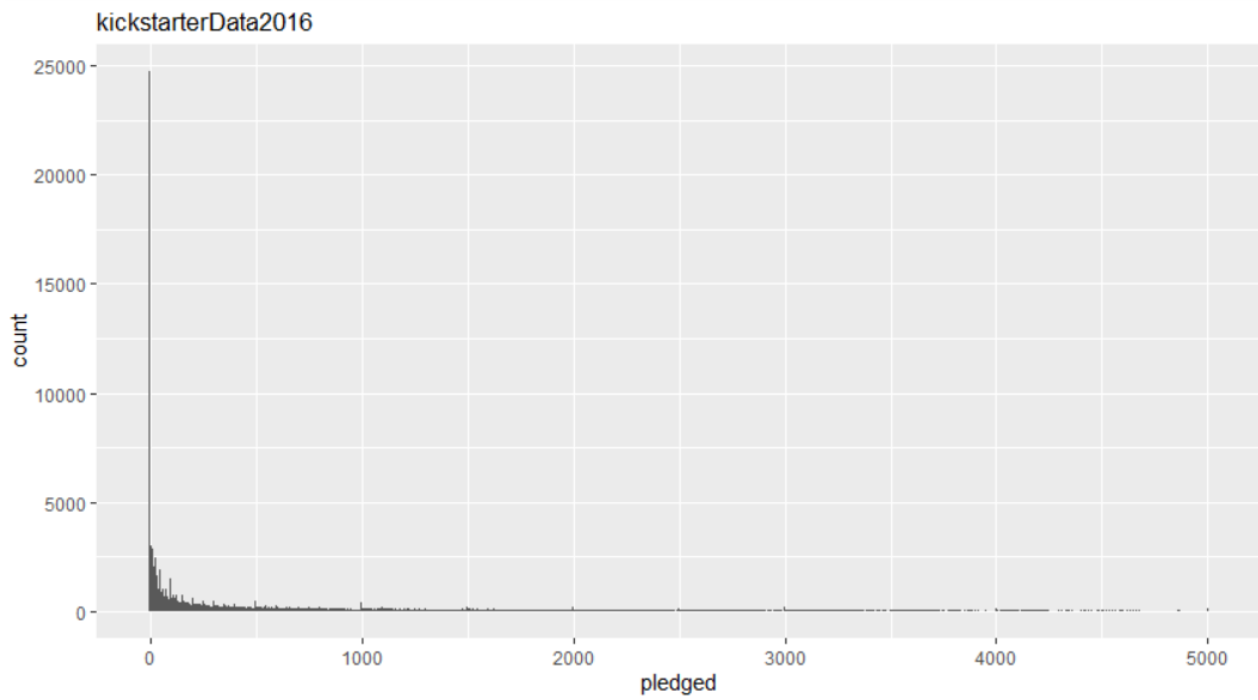
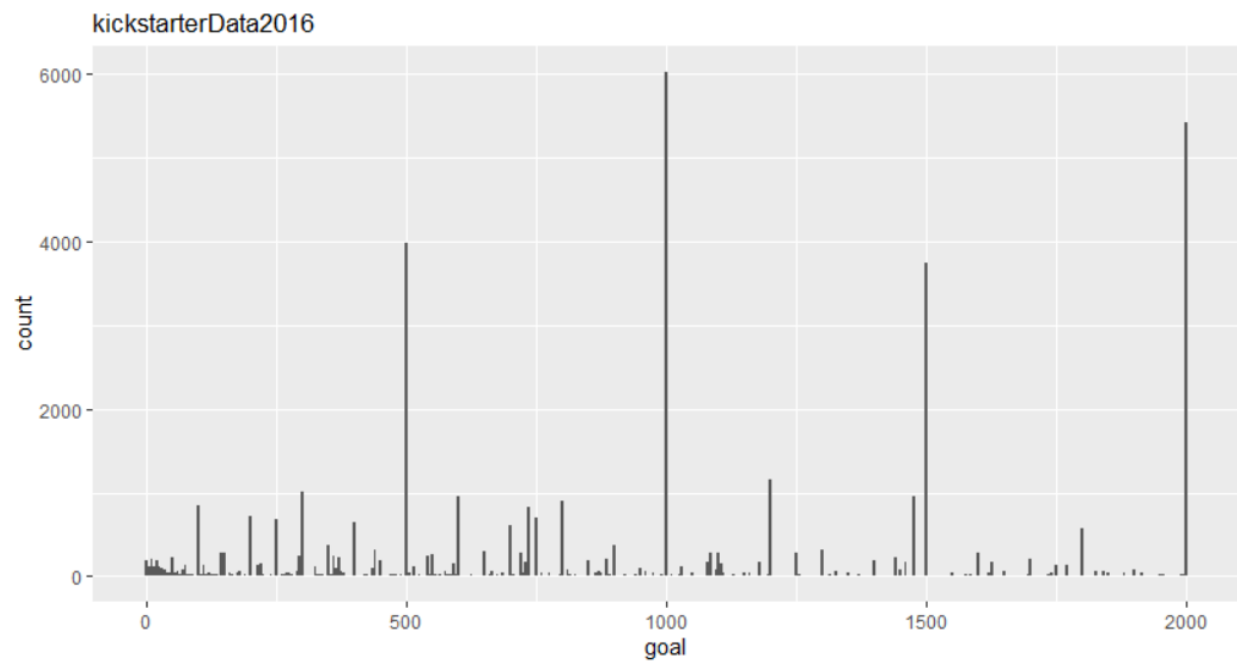


Distribúcia numerických atribútov. Podľa tejto distribúcie vidíme, že existujú nejaké špecifické prípady alebo anomálie.

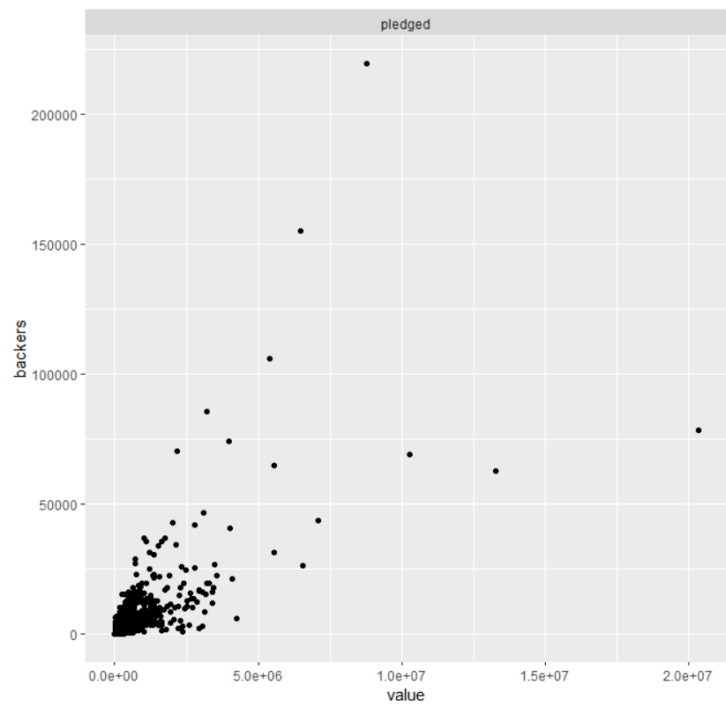


Vymazaním tých špecifických prípadov vidíme ako sa mení distribúcia.

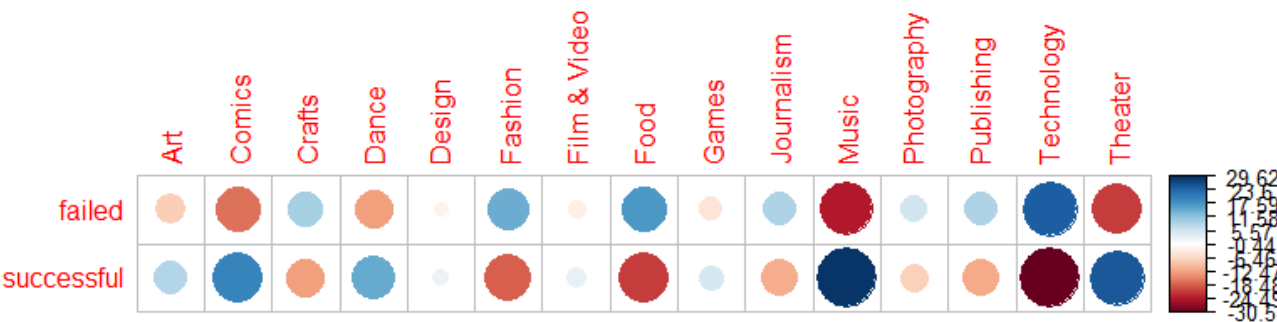




Rozptyl dát. Investované peniaze podľa počtu podporovateľov.



Chi-kvadrat test nezávislosti atribútov



Príprava dát

Stanovenie cieľa dát:

Na základne analýzy dát sme stanovili klasifikačný cieľ, predpovedať či daný projekt bude úspešný alebo nie.

Výber dát №1:

V tabuľke kickstarter 2016 sme vymazali stĺpce: X, X.1, X.2, X.3, name, deadline, launched, category, usd.pledged.

V tabuľke kickstarter 2018 sme vymazali stĺpce: name, deadline, launched, category, usd.pledged, usd_pledged_real, usd_goal_real.

Keďže pre ďalšie kroky dané stĺpce nepotrebujeme.

Filtrácia dát:

Keďže atribút state má viac hodnôt spolu s nekonzistentnými dátami, odfiltrovali sme ho na dve možné hodnoty, buď úspešný projekt alebo neúspešný("successful"/"failed").

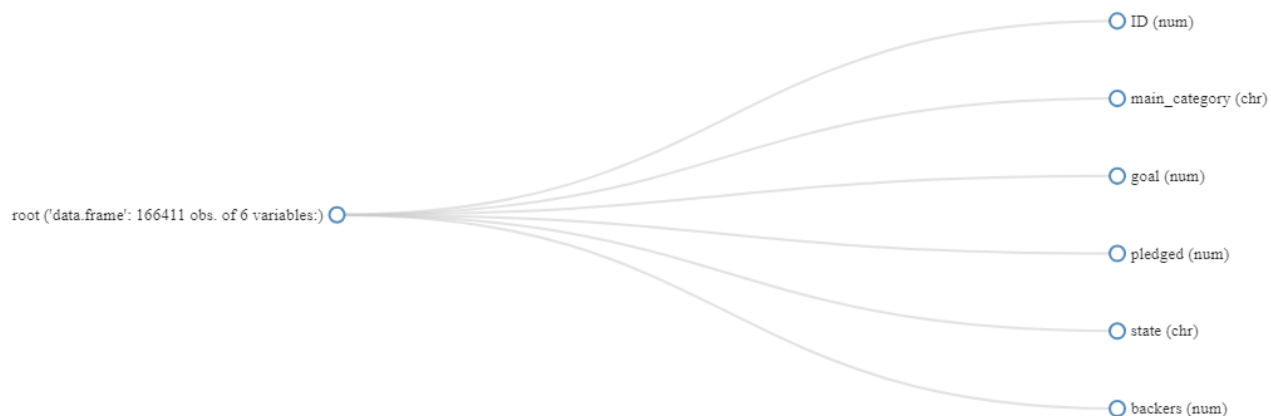
Konverzia mien:

Podľa výsledkov analýzy sme zistili, že uprostred mien prevažuje mena "USD". Teda všetky ostatné meny sme konvertovali na "USD".

Výber dát №2:

Po hore spomínaných úpravách sme vymazali v oboch tabuľkách stĺpce: currency a country.

Po úpravách tabuľky vhodné na ďalšiu analýzu, spojenie a vyzerajú nasledovné:



Spojenie tabuliek:

Výsledná a pripravená na splnenie klasifikačnej úlohy tabuľka vyzerá nasledovne:



Ako už bolo povedané, klasifikačnou úlohou týchto dát je predpovedať, či bude daný crowdfundingový projekt úspešný alebo nie.

Zmeny v príprave dát pre popisne úlohy

Filtrácia dát:

Odfiltrovali sme atribút state na štyri možné hodnoty, buď úspešný projekt, neúspešný projekt, prebiehajúci projekt alebo zrušený projekt ("successful"/"failed"/"live"/"canceled").

Výber dát №2:

Vymazali sme v oboch tabuľkách stĺpce: currency a ID.

Spojenie tabuliek:

Spojenie tabuliek pre popisne úlohy nebolo vykonané, cieľom totiž bolo porovnať dáta z roku 2016 a 2018.

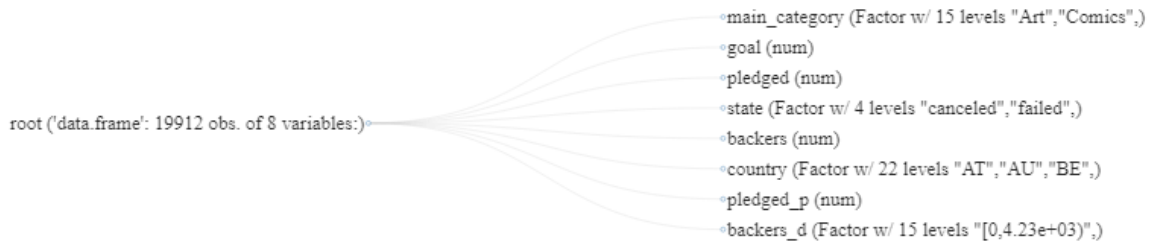
Na účel splnenia popisnej úlohy pomocou asociačných pravidiel sme diskretizovali atribút a "backers" na 15 intervalov, keďže podporovateľov bolo veľa, a previedli na percentuálnu hodnotu

atribút “pledged” podľa atribútu “goal”, teda koľko percenta od cieľového súčtu bolo investované do projektu.

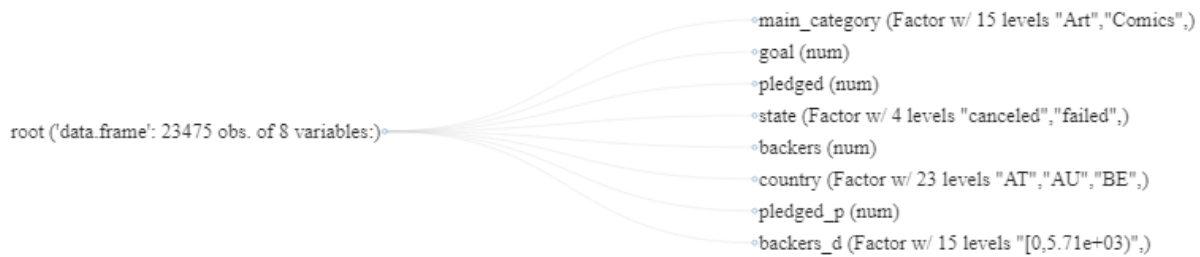
Na

Tabuľky:

Výsledná a pripravená na splnenie popisnej úlohy tabuľka pre rok 2016 vyzerá nasledovne:



Výsledná a pripravená na splnenie popisnej úlohy tabuľka pre rok 2018 vyzerá nasledovne:



Klasifikačná úloha

Rozdelenie dát na trénovaciu a testovaciu množinu:

Dáta boli rozdelené proporčne, teda 75% dát je trénovacia množina, keď 25% je testovacia.

Zvolené algoritmy na splnenie klasifikačnej úlohy:

- Rozhodovací strom
- Náhodný les
- Naivný Bayesovský klasifikátor

Vyhodnotenie výsledkov modelov:

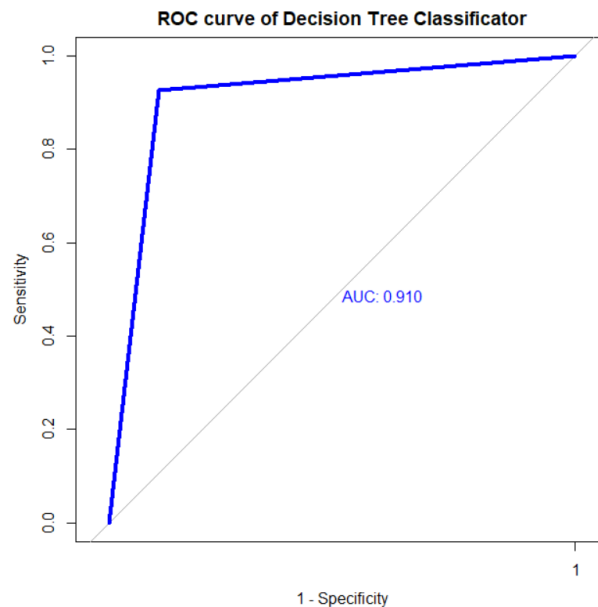
1. Rozhodovací strom:

Precision: 95%,

Recall: 89%,

Specificity: 85%,

ROC/AUC:



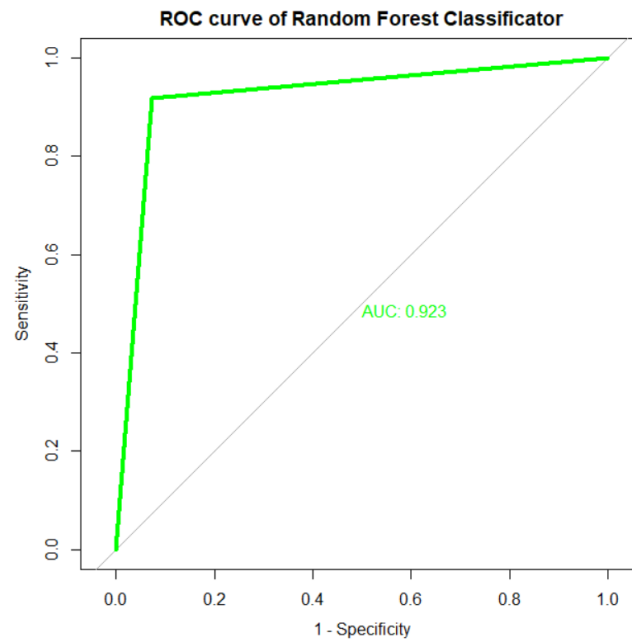
2. Náhodný les:

Precision: 94%,

Recall: 92%,

Specificity: 89%,

ROC/AUC:



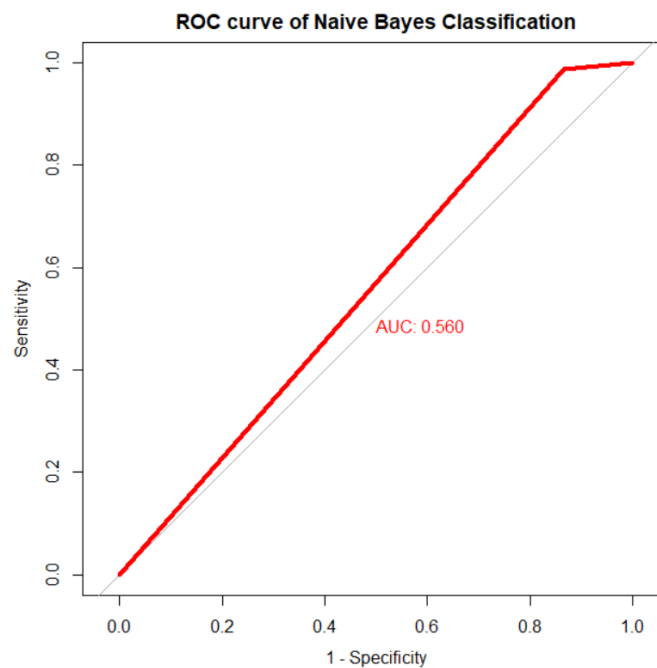
3. Naivný Bayesovský klasifikátor:

Precision: 94%,

Recall: 13%,

Specificity: 43%,

ROC/AUC:



Popisná úloha

Zvolené algoritmy na splnenie popisnej úlohy:

- Asociačné pravidla
- Zhhlukovanie

Vyhodnotenie výsledkov modelov:

1. Asociačné pravidla:

Pomocou asociačných pravidiel sme hľadali pravidla ktoré vedu k jednotlivým hodnotám pre atribút state

Kickstarter2016:

lhs	rhs	support	confidence	coverage	lift	count
[1] {main_category=Film & video, country=US, pledged_p=[101,1.7e+06]}	=> {state=successful}	0.05182804	0.9971014	0.05197871	2.777211	1032
[2] {main_category=Film & video, country=US, pledged_p=[101,1.7e+06], backers_d=[0,4.23e+03]}	=> {state=successful}	0.05177782	0.9970986	0.05192849	2.777204	1031
[3] {main_category=Film & video, pledged_p=[101,1.7e+06]}	=> {state=successful}	0.06066693	0.9958780	0.06091804	2.773804	1208
[4] {main_category=Film & video, pledged_p=[101,1.7e+06], backers_d=[0,4.23e+03]}	=> {state=successful}	0.06056649	0.9958712	0.06081760	2.773785	1206
[5] {main_category=Music, country=US, pledged_p=[101,1.7e+06]}	=> {state=successful}	0.05916031	0.9949324	0.05946163	2.771170	1178
[6] {main_category=Music, country=US, pledged_p=[101,1.7e+06], backers_d=[0,4.23e+03]}	=> {state=successful}	0.05916031	0.9949324	0.05946163	2.771170	1178
[7] {main_category=Music, pledged_p=[101,1.7e+06]}	=> {state=successful}	0.06578947	0.9931766	0.06624146	2.766280	1310
[8] {main_category=Music, pledged_p=[101,1.7e+06], backers_d=[0,4.23e+03]}	=> {state=successful}	0.06578947	0.9931766	0.06624146	2.766280	1310
[9] {country=US, pledged_p=[101,1.7e+06]}	=> {state=successful}	0.27952993	0.9926877	0.28158899	2.764918	5566
[10] {country=US, pledged_p=[101,1.7e+06], backers_d=[0,4.23e+03]}	=> {state=successful}	0.27807352	0.9926497	0.28013258	2.764812	5537
[11] {pledged_p=[101,1.7e+06]}	=> {state=successful}	0.33155886	0.9899535	0.33492366	2.757302	6602
[12] {pledged_p=[101,1.7e+06], backers_d=[0,4.23e+03]}	=> {state=successful}	0.32955002	0.9898929	0.33291483	2.757133	6562
[13] {country=US, pledged_p=[0,2]}	=> {state=failed}	0.20033146	0.8170832	0.24517879	1.556170	3989
[14] {country=US, pledged_p=[0,2], backers_d=[0,4.23e+03]}	=> {state=failed}	0.20033146	0.8170832	0.24517879	1.556170	3989
[15] {pledged_p=[0,2]}	=> {state=failed}	0.25939132	0.8113415	0.31970671	1.545235	5165
[16] {pledged_p=[0,2], backers_d=[0,4.23e+03]}	=> {state=failed}	0.25939132	0.8113415	0.31970671	1.545235	5165
[17] {country=US, pledged_p=[2,101]}	=> {state=failed}	0.21650261	0.7811198	0.27716955	1.487676	4311
[18] {country=US, pledged_p=[2,101], backers_d=[0,4.23e+03]}	=> {state=failed}	0.21650261	0.7811198	0.27716955	1.487676	4311
[19] {pledged_p=[2,101]}	=> {state=failed}	0.26566894	0.7692308	0.34536963	1.465033	5290
[20] {pledged_p=[2,101], backers_d=[0,4.23e+03]}	=> {state=failed}	0.26566894	0.7692308	0.34536963	1.465033	5290

Kickstarter2018:

lhs	rhs	support	confidence	coverage	lift	count
[1] {main_category=Music, country=US, pledged_p=[101,1.04e+07]}	=> {state=successful}	0.05363152	0.9976228	0.05375932	2.750992	1259
[2] {main_category=Music, country=US, pledged_p=[101,1.04e+07], backers_d=[0,5.71e+03]}	=> {state=successful}	0.05363152	0.9976228	0.05375932	2.750992	1259
[3] {main_category=Music, pledged_p=[101,1.04e+07]}	=> {state=successful}	0.05989350	0.9971631	0.06006390	2.749724	1406
[4] {main_category=Music, pledged_p=[101,1.04e+07], backers_d=[0,5.71e+03]}	=> {state=successful}	0.05989350	0.9971631	0.06006390	2.749724	1406
[5] {main_category=Film & video, pledged_p=[101,1.04e+07]}	=> {state=successful}	0.05861555	0.9971014	0.05878594	2.749554	1376
[6] {main_category=Film & video, pledged_p=[101,1.04e+07], backers_d=[0,5.71e+03]}	=> {state=successful}	0.05848775	0.9970951	0.05865815	2.749537	1373
[7] {country=US, pledged_p=[101,1.04e+07]}	=> {state=successful}	0.27382322	0.9925880	0.27586794	2.737108	6428
[8] {country=US, pledged_p=[101,1.04e+07], backers_d=[0,5.71e+03]}	=> {state=successful}	0.27254526	0.9925535	0.27458999	2.737013	6398
[9] {pledged_p=[101,1.04e+07]}	=> {state=successful}	0.33669862	0.9917189	0.33951012	2.734712	7904
[10] {pledged_p=[101,1.04e+07], backers_d=[0,5.71e+03]}	=> {state=successful}	0.33495208	0.9916761	0.33776358	2.734594	7863
[11] {country=US, pledged_p=[0,2]}	=> {state=failed}	0.19446219	0.8251988	0.23565495	1.561088	4565
[12] {country=US, pledged_p=[0,2], backers_d=[0,5.71e+03]}	=> {state=failed}	0.19446219	0.8251988	0.23565495	1.561088	4565
[13] {pledged_p=[0,2]}	=> {state=failed}	0.25857295	0.8170682	0.31646432	1.545707	6070
[14] {pledged_p=[0,2], backers_d=[0,5.71e+03]}	=> {state=failed}	0.25857295	0.8170682	0.31646432	1.545707	6070
[15] {country=US, pledged_p=[2,101]}	=> {state=failed}	0.21329073	0.7916206	0.26943557	1.497566	5007
[16] {country=US, pledged_p=[2,101], backers_d=[0,5.71e+03]}	=> {state=failed}	0.21329073	0.7916206	0.26943557	1.497566	5007
[17] {pledged_p=[2,101]}	=> {state=failed}	0.27003195	0.7849183	0.34402556	1.484886	6339
[18] {pledged_p=[2,101], backers_d=[0,5.71e+03]}	=> {state=failed}	0.27003195	0.7849183	0.34402556	1.484886	6339

Pomocou asociačných pravidiel sme pre obe tabuľky dokázali nájsť pravidla len pre hodnotu “successful” a “failed”.

Keďže hodnota “live” stále nedošla do koncového stavu obsahuje dáta ktoré sa ešte môžu zmeniť a pomocou asociačných pravidiel je problematické priradiť tomuto stavu pravidla pre veľmi rôznorodé hodnoty.

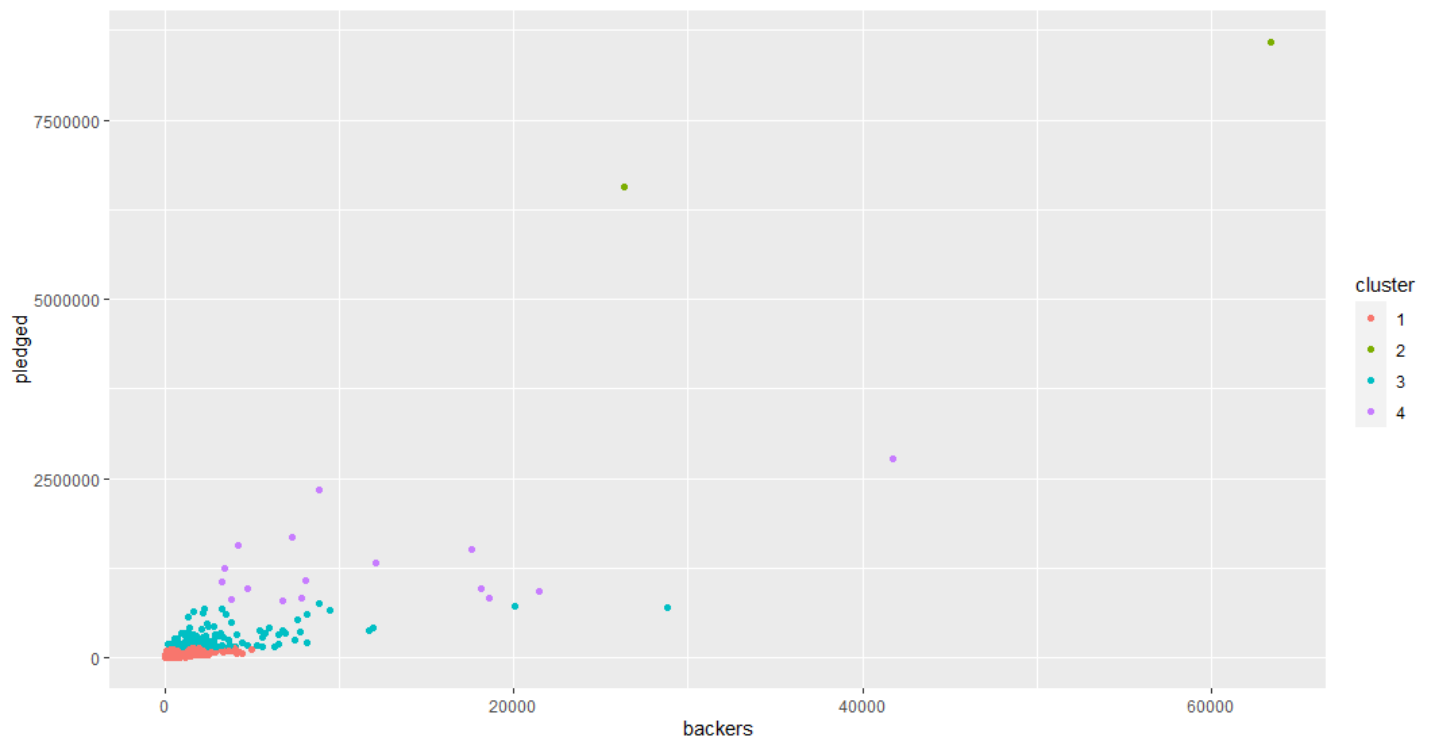
Hodnoty “live” a “canceled” obsahujú menšie množstvo dát ako hodnoty “successful” a “failed”, preto je problematickejšie pre nich vytvoriť pravidla.

2. Zhhlukovanie

Pre zhhlukovanie sme použili atribúty pledged, backers. Tieto atribúty sme zhľukovali do 4 tried.

Kickstarter2016:

Vizualizácia zhukovania podľa atribútov pledged a backers



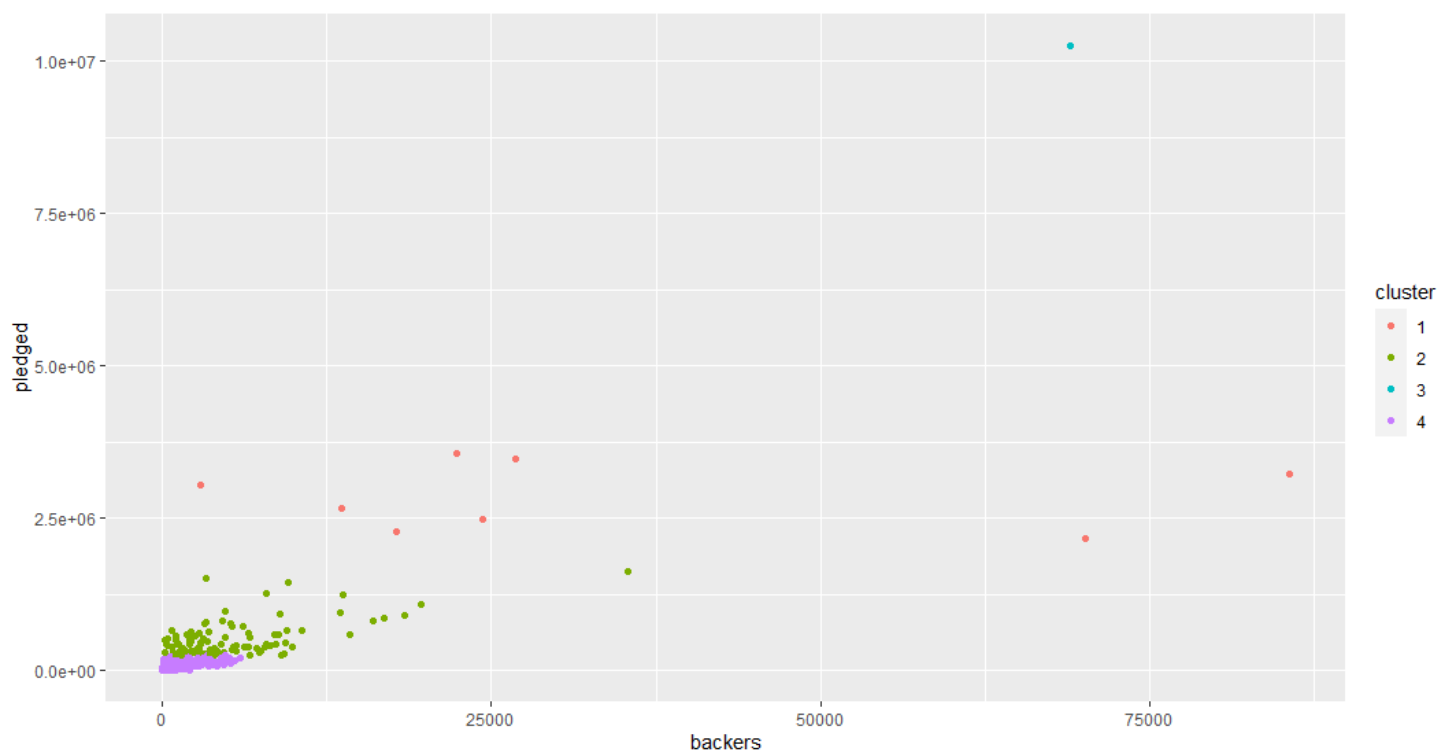
Vyhodnotenie zhukovania:

Čistota zhukov: 0.53

Gini index: 0.0056

Kickstarter2018:

Vizualizácia zhukovania podľa atribútov pledged a backers



Vyhodnotenie zhlučovania:

Čistota zhlukov: 0.53

Gini index: 0.0089

Vyhodnotenie

Prediktívne Modely:

Na tréovanie modelov boli použité atribúty: ID, Goal, Backers a State

Model	Precision	Recall	Specificity	AUC
Rozhodovací strom	95%	89%	85%	0.910
Náhodný les	94%	92%	98%	0.923
Naivný Bayesovský klasifikátor	94%	13%	43%	0.560

Na riešenie klasifikačnej úlohy najvhodnejším modelom je Náhodný les (podľa výsledných metrík), ďalší model je Rozhodovací strom, ktorý dosiahol veľmi uspokojujúce výsledky.

Model Naivného Bayesovského klasifikátora je tiež použiteľný, avšak, nemá až také vyhovujúce výsledky ako rozhodovací strom a náhodný les a podľa výsledkov si môžeme všimnúť, že je citlivý na nerovnomernú distribúciu dát, čo hovoria vysoká presnosť a nízka návratnosť, preto by pri riešení reálneho problému nebol využitý.

Podľa výsledkov daných modelom môžeme zhrnúť, že najvhodnejšími modelmi pre tieto dáta by boli algoritmy na základe rozhodovacích stromov, keďže nepotrebujú veľmi starostlivého predspracovania, škálovania dát, rovnomernej distribúcie atď. Naivný Bayesovský klasifikátor takú výhodu však nemá.

Popisne Modely:

Pre zhlučovanie boli použité atribúty : pledged a backers

Pri zhlučovaní bola použitá metóda K-means zhlučovania do 4 tried.

Zhlučovanie bolo pre obe tabuľky veľmi podobne, čistota zhlukov bola pre obe úplne rovnaká a hodnota Gini index bola pre tabuľky veľmi dobrá kde o niečo lepšia sú zhluky pre Kickstarter2016. Avšak rozdiel je skoro zanedbateľný s rozdielom len 0.0033

Tabuľka	Čistota zhlukov	Gini index
Kickstarter2016	0.53	0.0056
Kickstarter2018	0.53	0.0089

Pomocou K-means zhlučovania nie je možné rozdeliť dáta do 4 tried keďže hodnoty sa prekrývajú a nie je možné ich separovať a na túto úlohu by mohlo byť vhodné zhlučovanie podľa hustoty (napr. DBSCAN).

(DBSCAN nebolo možné použiť pre zahltenie pamäte.)

Pre asociačné pravidla boli použité atribúty: main_category, state, country, pledged_p, backers_d.

Na vytvorenie asociačných pravidiel sme zvolili nasledujúce parametre: support(podpora) 0.05, confidence(spoľahlivosť) 0.7 a minimálnu dĺžku pravidiel 2.

Zaujímavé pravidla:

1. Rok 2016

lhs	rhs	support	confidence	coverage	lift	count
[1] {main_category=Film & video, country=US, pledged_p=[101,1.7e+06]}	=> {state=successful}	0.05182804	0.9971014	0.05197871	2.777211	1032
[2] {main_category=Film & video, country=US, pledged_p=[101,1.7e+06], backers_d=[0,4.23e+03]}	=> {state=successful}	0.05177782	0.9970986	0.05192849	2.777204	1031
[3] {main_category=Film & video, pledged_p=[101,1.7e+06]}	=> {state=successful}	0.06066693	0.9958780	0.06091804	2.773804	1208
[4] {main_category=Film & video, pledged_p=[101,1.7e+06], backers_d=[0,4.23e+03]}	=> {state=successful}	0.06036649	0.9958712	0.06081760	2.773785	1206

Podľa pravidiel s odstavca o popisnej úlohy si môžeme všimnúť, že v úspešných projektoch v roku 2016 prevažuje väčšinou kategória "Film&Video", a presné táto kategória má najviac podporovateľov. To, že daná kategória je najviac populárna hovoria aj barploty s distribúciou kategórií v dátovej množine. Prvé štyri pravidla, vygenerované pre rok 2016, ktoré majú najväčšiu spoľahlivosť, hovoria nám o približnom súčte, ktorý robil projekt úspešným. Taktiež vidíme, že do kategórie "Film&Video" bolo investované väčšinu peniazi, čo zas potvrdzuje popularitu danej kategórie.

[9] {country=US, pledged_p=[101,1.7e+06]}	=> {state=successful}	0.27952993	0.9926877	0.28158899	2.764918	5566
[10] {country=US, pledged_p=[101,1.7e+06], backers_d=[0,4.23e+03]}	=> {state=successful}	0.27807352	0.9926497	0.28013258	2.764812	5537
[11] {pledged_p=[101,1.7e+06]}	=> {state=successful}	0.33155886	0.9899535	0.33492366	2.757302	6602

Hore uvedené pravidla tiež potvrdzujú približný súčet, ktorý je na 98% zárukou, že projekt, do ktorého bolo investované peniazi z uvedeného v pravidlách intervalu, bude úspešný.

[5] {main_category=Music, country=US, pledged_p=[101,1.7e+06]}	=> {state=successful}	0.05916031	0.9949324	0.05946163	2.771170	1178
[6] {main_category=Music, country=US, pledged_p=[101,1.7e+06], backers_d=[0,4.23e+03]}	=> {state=successful}	0.05916031	0.9949324	0.05946163	2.771170	1178
[7] {main_category=Music, pledged_p=[101,1.7e+06]}	=> {state=successful}	0.06578947	0.9931766	0.06624146	2.766280	1310
[8] {main_category=Music, pledged_p=[101,1.7e+06], backers_d=[0,4.23e+03]}	=> {state=successful}	0.06578947	0.9931766	0.06624146	2.766280	1310

Taktiež podľa pravidiel si môžeme všimnúť, že podľa popularity druhou kategóriou je Hudba.

2. Rok 2018

lhs	rhs	support	confidence	coverage	lift	count
[1] {main_category=Music, country=US, pledged_p=[101,1.04e+07]}	=> {state=successful}	0.05363152	0.9976228	0.05375932	2.750992	1259
[2] {main_category=Music, country=US, pledged_p=[101,1.04e+07], backers_d=[0,5.71e+03]}	=> {state=successful}	0.05363152	0.9976228	0.05375932	2.750992	1259
[3] {main_category=Music, pledged_p=[101,1.04e+07]}	=> {state=successful}	0.05989350	0.9971631	0.06006390	2.749724	1406
[4] {main_category=Music, pledged_p=[101,1.04e+07], backers_d=[0,5.71e+03]}	=> {state=successful}	0.05989350	0.9971631	0.06006390	2.749724	1406
[5] {main_category=Film & video, pledged_p=[101,1.04e+07]}	=> {state=successful}	0.05861555	0.9971014	0.05878594	2.749554	1376
[6] {main_category=Film & video, pledged_p=[101,1.04e+07], backers_d=[0,5.71e+03]}	=> {state=successful}	0.05848775	0.9970951	0.05865815	2.749537	1373

V roku 2018 môžeme vidieť, že tak isto ako aj v roku 2016 najpopulárnejšími kategóriami sú "Music" a "Film&Video". Tiež si môžeme všimnúť súčet, ktorý je vhodný na to, aby projekt bol úspešným, aj keď sa trochu líši od roku 2016, aj tak globálne je približné rovnaký. Počet podporovateľov však nie je 100% zárukou, že projekt bude úspešný, čo znázorňujú pravidla hore ako aj nasledujúce.

[17] {pledged_p=[2,101]}	=> {state=failed}	0.27003195	0.7849183	0.34402556	1.484886	6339
[18] {pledged_p=[2,101], backers_d=[0,5.71e+03]}	=> {state=failed}	0.27003195	0.7849183	0.34402556	1.484886	6339

Asociačné pravidla sú veľmi užitočnou popisnou metódou, ktorá pomáha vyjadriť asociácie atribútov a zistiť závislosť hodnôt cieľového atribútu od hodnôt ostatných atribútov. V našom prípade to nám pomohlo zistiť približný súčet na to, aby viac projektov s rôznym cieľovým súčtom bolo úspešný.