I-SUNS: Zadanie č.2

Analýa dát a regresory – JupyterLab Python

Analýza

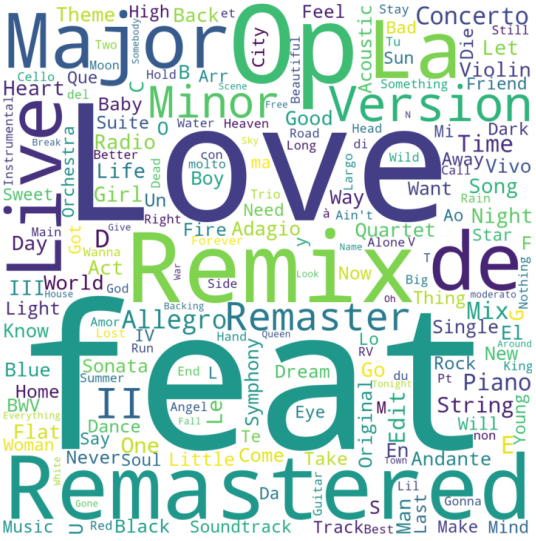
Ako prvé som si načítal dáta, ktoré som pre analýzu spojil do jedného celku. Zisťoval som, koľko originálnych žánrov sa nachádza v stĺpci *artist_genres* – 3130 alebo taktiež, ktoré žánre boli najviac bežné. Ktorí hudobníci boli najviac vyskytujúci.

[('rock', 6316),
('pop', 5089),
('dance pop', 3960),
('classic rock', 3412),
('classical', 3253),
('rap', 2177),
('modern rock', 2139),
('mellow gold', 2087),
('hip hop', 1979),
('album rock', 1954),
('permanent wave', 1833),
('alternative metal', 1773),
('edm', 1743),
('pop rap', 1713),
('soft rock', 1683),
('hard rock', 1640),
('alternative rock', 1598),
('pop rock', 1551),
('r&b', 1489),
('urban contemporary', 1386)]

[('Queen', 531),
('Wolfgang Amadeus Mozart', 240),
('Ludwig van Beethoven', 227),
('Johann Sebastian Bach', 216),
('Taylor Swift', 186),
('The Beatles', 184),
('Status Quo', 179),
('Vitamin String Quartet', 153),
('Metallica', 145),
('Nicki Minaj', 131),
('Antonio Vivaldi', 130),
('Claude Debussy', 130),
('Selena', 120),
('Aretha Franklin', 110),
('Kanye West', 108),
('John Williams', 106),
('Franz Schubert', 104),
('Erik Satie', 102),
('Nick Neblo', 100),
('Madonna', 99)]

Top 20 najviac vyskytujúcich sa žánrov *Top 20 najviac vyskytujúcich sa hudobníkov*

Medzi ďalšie veci patril WordCloud pre najčastejšie sa vyskytujúce sa slová v názvoch pesničiek.



Dávid Gavenda 98533

Nasledovalo zisťovanie informácií o našom DataFrame. Kde som zisťoval aké premenné sa tam nachádzajú a ich unikátne počty. Priemery, rozptyl ako aj rozdelenie do kvartilov. *Mode* a *explicit* majú len dve hodnoty, teda môžu byť zastúpené čisto binárne, *explicit* má ale bool, takže to zmeníme.

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   id                53669 non-null  object          53669
 1   artist_id         53669 non-null  object          14034
 2   artist            53669 non-null  object          14008
 3   name              53669 non-null  object          45074
 4   popularity        53669 non-null  int64           100
 5   release_date      53669 non-null  object          7216
 6   duration_ms       53669 non-null  int64           32738
 7   explicit          53669 non-null  bool            2
 8   danceability      53669 non-null  float64         1161
 9   energy            53669 non-null  float64         2193
 10  key               53669 non-null  int64           12
 11  loudness          53669 non-null  float64         18363
 12  mode              53669 non-null  int64           2
 13  speechiness       53669 non-null  float64         1312
 14  acousticness      53669 non-null  float64         4445
 15  instrumentalness  53669 non-null  float64         5242
 16  liveness          53669 non-null  float64         1684
 17  valence           53669 non-null  float64         1636
 18  tempo             53669 non-null  float64         34458
 19  artist_genres     53669 non-null  object          9414
 20  artist_followers  53668 non-null  float64         13283
 21  url               53669 non-null  object          53669
 22  playlist_id       53669 non-null  object          1046
 23  playlist_description  37150 non-null  object      722
 24  playlist_name     53643 non-null  object          1027
 25  playlist_url      53669 non-null  object          1046
 26  query             53669 non-null  object          17
```

Neskôr potrebné ošetriť

	popularity	duration_ms	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	artist_followers
count	53669.00	53669.00	53669.00	53669.00	53669.00	53669.00	53669.00	53669.00	53669.00	53669.00	53669.00	53669.00	53669.00	53668.00
mean	35.38	241066.19	0.55	0.60	5.25	-9.38	0.62	0.08	0.32	0.19	0.18	0.45	121.21	3050833.05
std	23.62	102419.53	0.19	0.28	3.56	6.38	0.49	0.09	0.35	0.33	0.16	0.26	30.83	7878787.32
min	0.00	11173.00	0.00	0.00	0.00	-44.94	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
25%	15.00	187013.00	0.43	0.41	2.00	-11.16	0.00	0.04	0.02	0.00	0.09	0.23	96.21	47032.00
50%	39.00	224306.00	0.56	0.65	5.00	-7.34	1.00	0.05	0.16	0.00	0.12	0.44	120.05	394103.00
75%	54.00	271626.00	0.69	0.83	8.00	-5.26	1.00	0.09	0.61	0.21	0.22	0.66	140.08	2140548.00
max	100.00	3590693.00	0.99	1.00	11.00	3.11	1.00	0.95	1.00	1.00	1.00	1.00	243.03	86319453.00

Odstránil som riadky v trénovacích dátach, kde sa *artist_id*, *name* a *duration_ms* zhodovali a taktiež vyhodil stĺpce, ktoré neboli potrebné pre náš účel, keďže neobsahovali číselné hodnoty ani nič užitočné.

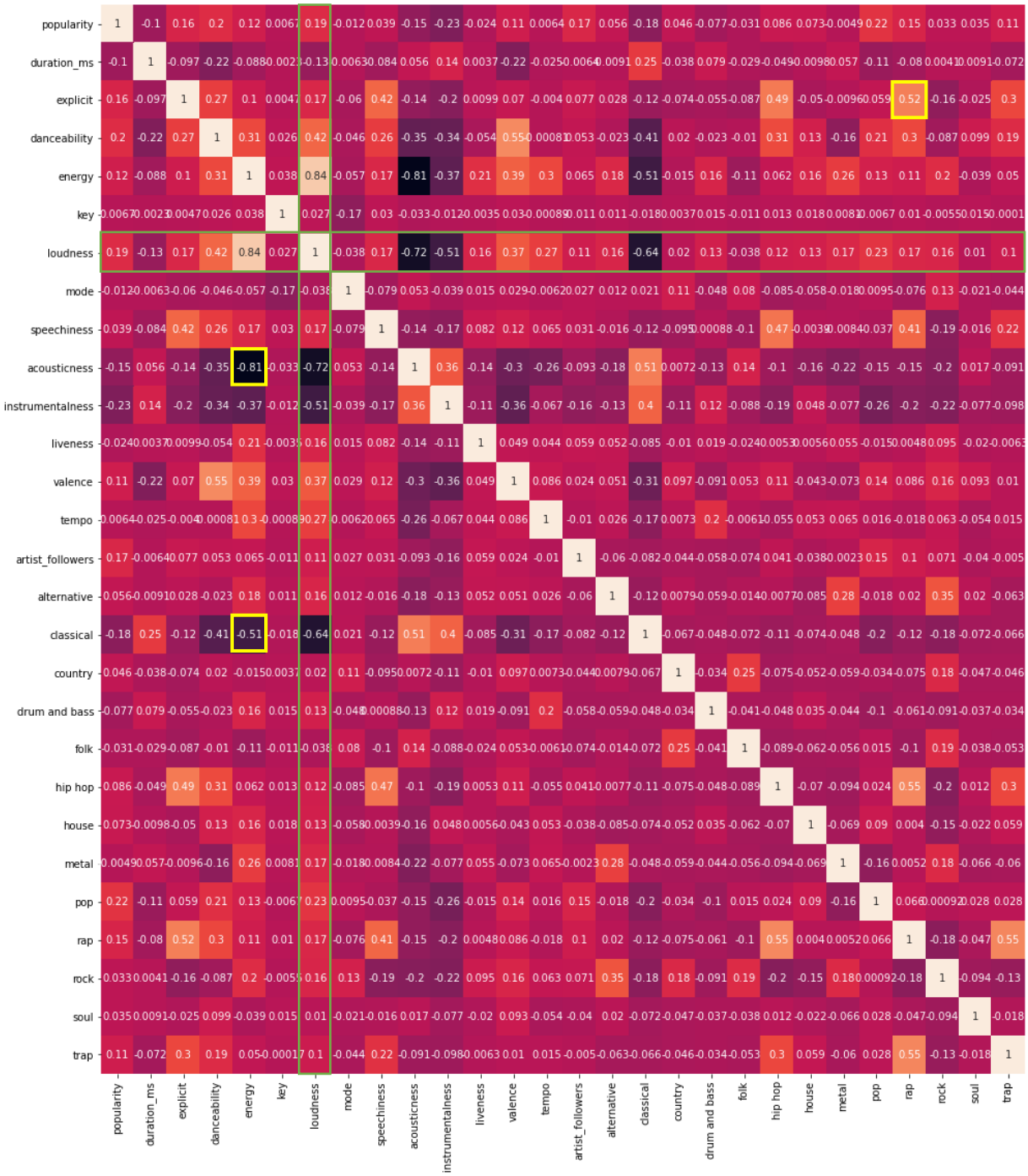
['artist', 'name','id','artist_id','url','playlist_id','playlist_description','playlist_name','playlist_url','query']

*po spracovaní aj *artist_genres*

Nahradil som hodnoty v *explicit* za binárne hodnoty, rovnako aj spracoval žánre, ktoré som rozdelil do 13 skupín ('alternative','classical','country','drum and bass','folk','hip hop','house', 'metal', 'pop', 'rap',

'rock', 'soul', 'trap'). Keďže žánrov je tam 3130, toto rozdelenie je omnoho viac efektívne, nemusí byť nutne presnejšie ale výsledok je dostatočný.

Nasledovala kovariačná matica, na základe ktorej sme sa dozvedeli, že *energy* a *loudness* spolu vysoko súvisia (0.84) alebo že *accousticness* s *loudness* naopak len minimálne (-0.72). Hodnoty mali samy so sebou 1, čo je vidno na diagonále matice. Stĺpec vyznačený je ten, ktorý nás zaujíma (analogicky môžeme zvoliť aj riadok). Teda hodnoty, ktoré nás najviac zaujali sú ***danceability (0.42), energy(0.84), valence(0.37)*** a ***acousticness(-0.72), instrumentalness(-0.51)*** a ***classical (-0.64).***



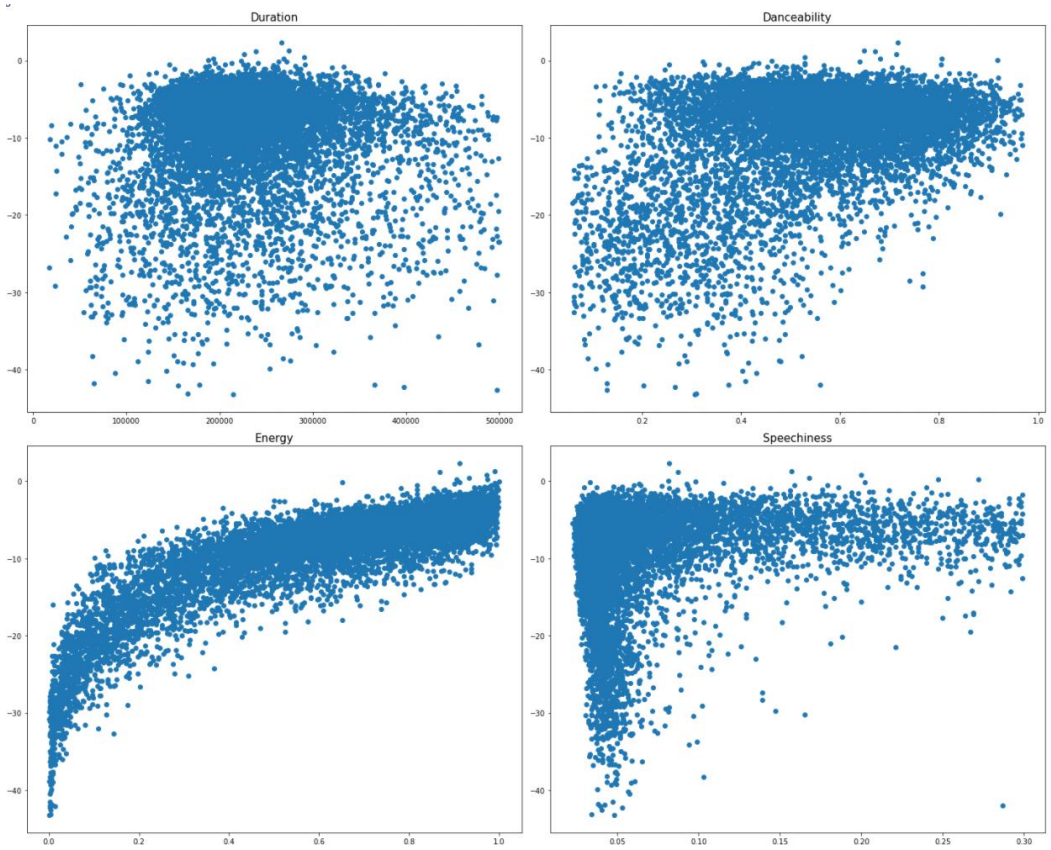
Dávid Gavenda 98533

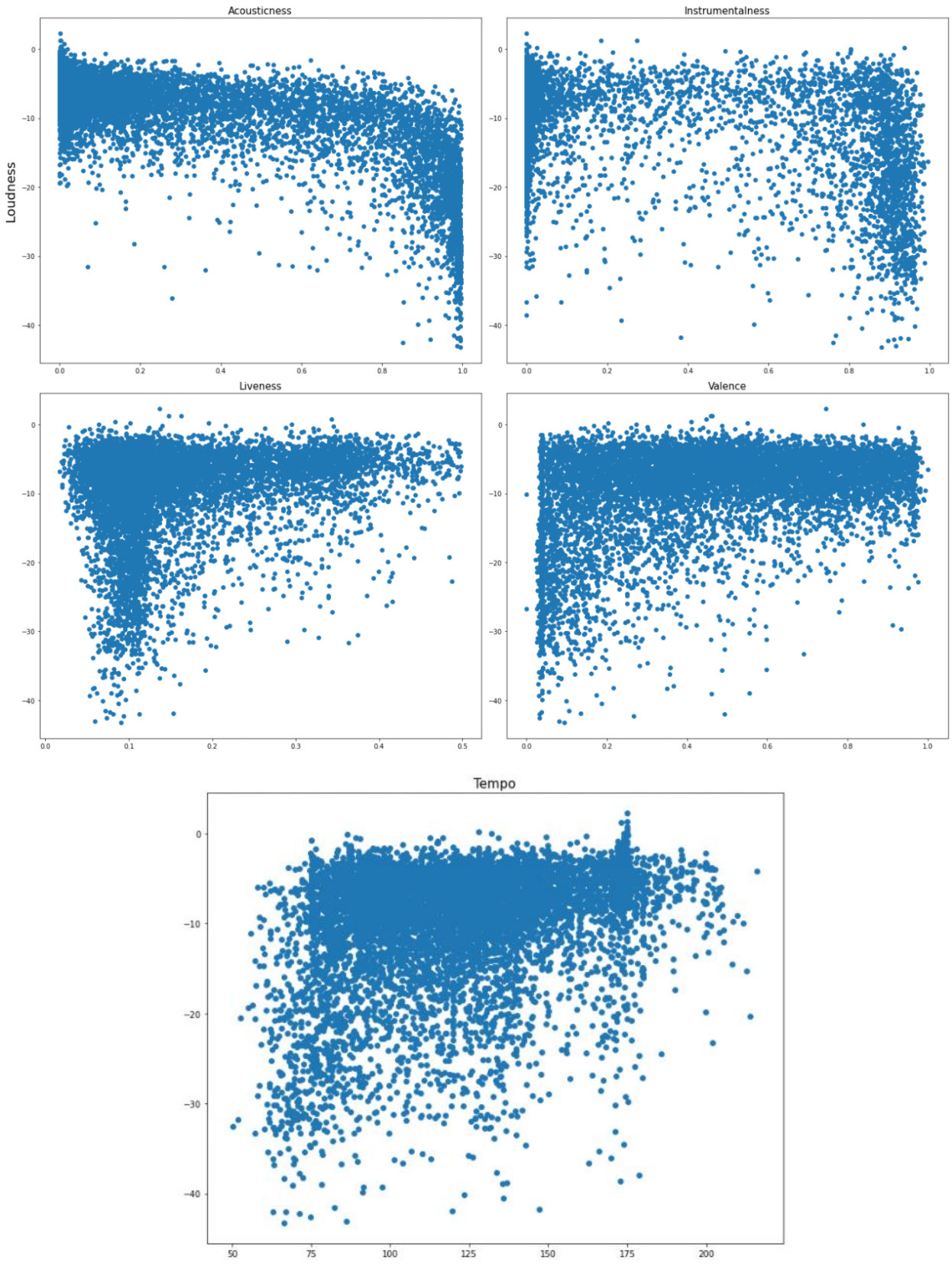
Väčšinou nadávky boli v rape, acousticness moc nejde s energy, tak ako ani classical.

Vypísal som si TOP 5 a BOTTOM 5 v *loudness s ostatnými stĺpcami, ktoré mali tiež vysoký vplyv. Všimneme si, že stĺpce, ktoré mali kladnú hodnotu (modrá) v matici, majú vyššie hodnoty, pri „hlučnejších" pesničkách a naopak, tie ktoré mali zápornú (žltá) nižšie hodnoty. V prípade, že by sme chceli dosiahnuť nižšiu záťaž na PC, taktiež časovú. Jednalo by sa o ideálnych kandidátov.*

	artist	name	loudness	energy	danceability	valence	acousticness	instrumentalness	classical
6940	Macky Gee	Games	3.108	0.99300	0.5910	0.1950	0.20900	0.049400	0
51289	Subsonic	Do Your Thang	2.292	0.91200	0.7170	0.7460	0.00109	0.000000	0
30055	Macky Gee	Hoe Talk	2.124	0.99000	0.6640	0.1250	0.01360	0.226000	0
29711	Macky Gee	Try	2.001	0.98000	0.6800	0.6910	0.01950	0.000548	0
797	Levela	Lights	1.858	0.99800	0.6240	0.4750	0.00932	0.859000	0
...	...	...	...	...	...	...	...	...	...
16532	Esa-Pekka Salonen	Violin Concerto: Movement Two: Pulse I	-43.277	0.00219	0.0754	0.0335	0.41100	0.296000	1
8813	Maurice Ravel	Ma mère l'oye, M. 62: Pavane de la belle au bo...	-43.812	0.00123	0.0921	0.0385	0.82600	0.112000	1
9057	John Cage	Dream	-44.135	0.00154	0.2660	0.0922	0.99300	0.862000	1
43046	Thomas Adès	Adès: Traced Overhead, Op. 15: I. Sursum	-44.727	0.01500	0.2040	0.0548	0.97800	0.929000	1
26249	Jeremy Soule	Shattered Shields	-44.945	0.00793	0.1580	0.0368	0.95400	0.963000	0

Vykreslil si vzťah medzi ***loudness*** a ostatnými číselnými hodnotami okrem ***release_date*** *(z ktorého som si vybral len rok a uložil ako číselnú premennú)* a ***artist_genres*** (teraz už osobitne každý žáner)





Dávid Gavenda 98533

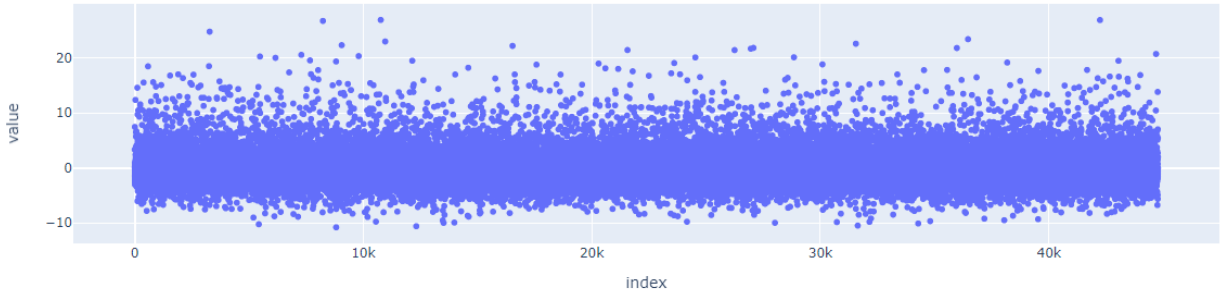
Pomocou zscore som následne odstránil **outliers** z trénovacích dát, približne 4% celkového počtu.

V neposlednom rade som **normalizoval** obe množiny.

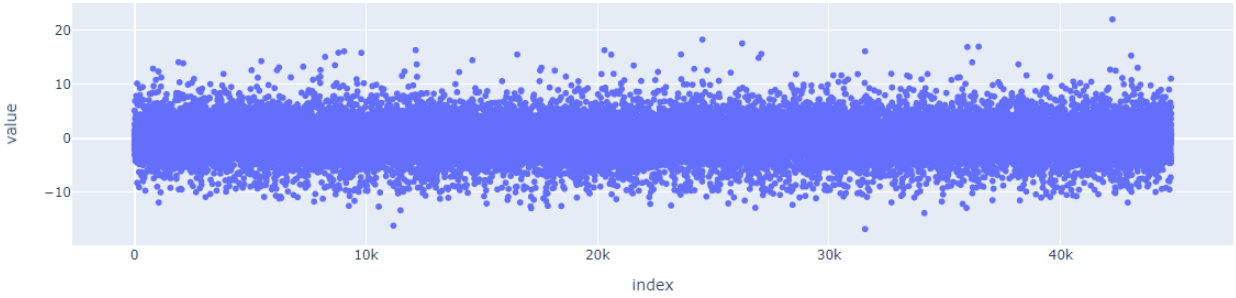
Regresory

Využil som 5 regresorov, ktorých výsledky sú zapísané v tabuľke, ako aj grafy reziduálov, ktoré sú nižšie. Pomocou SVM regresora a jeho defaultných nastavení, kde boli vstupom hodnoty, ktoré som spracoval okrem *loudness,* ktoré bolo v druhej množine.

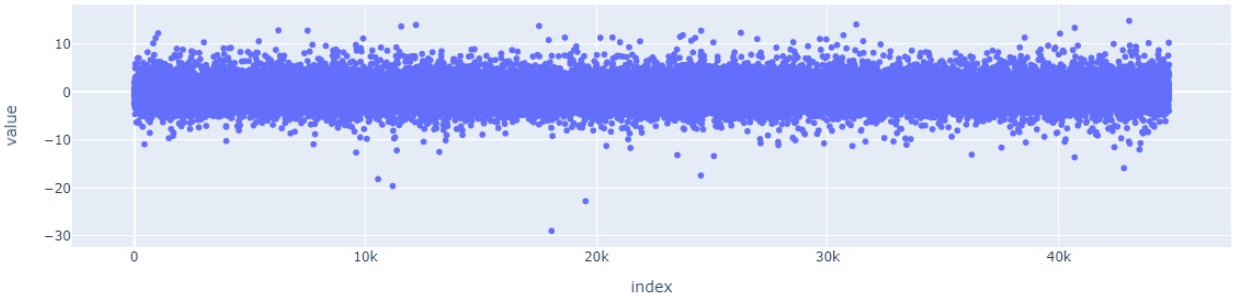
Názov	Score – R^2	MSE	Accuracy
SVM	0.7968826691268804	8.14730760787935	0.86
GridSearch	0.8218631538733672	7.1453069782496685	* na dlho
RandomForest	**0.8819663881408727**	**4.734485923734688**	**0.89**
Bagging	0.7964930864998184	8.162934287728456	0.86
Boosting	0.8817828798726488	4.7418466856321695	0.88



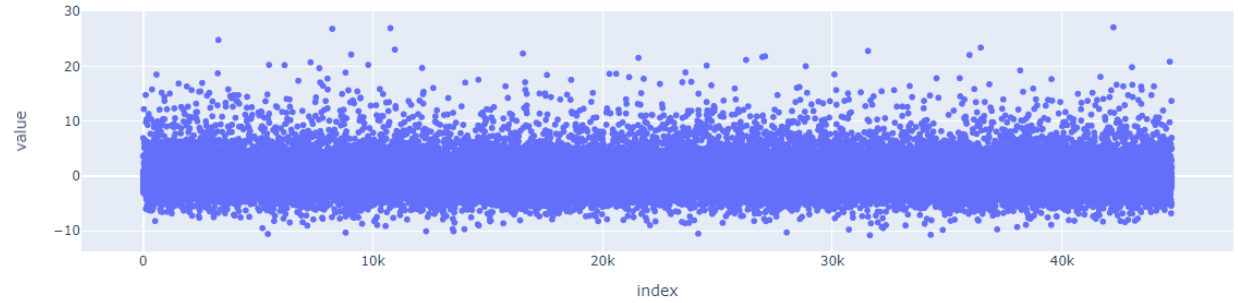
SVM - reziduály



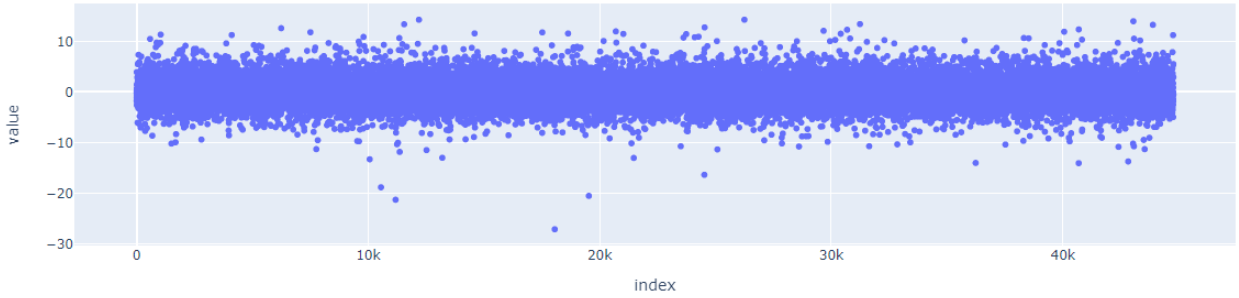
GridSearch - reziduály



RandomForest - reziduály



Bagging - reziduály

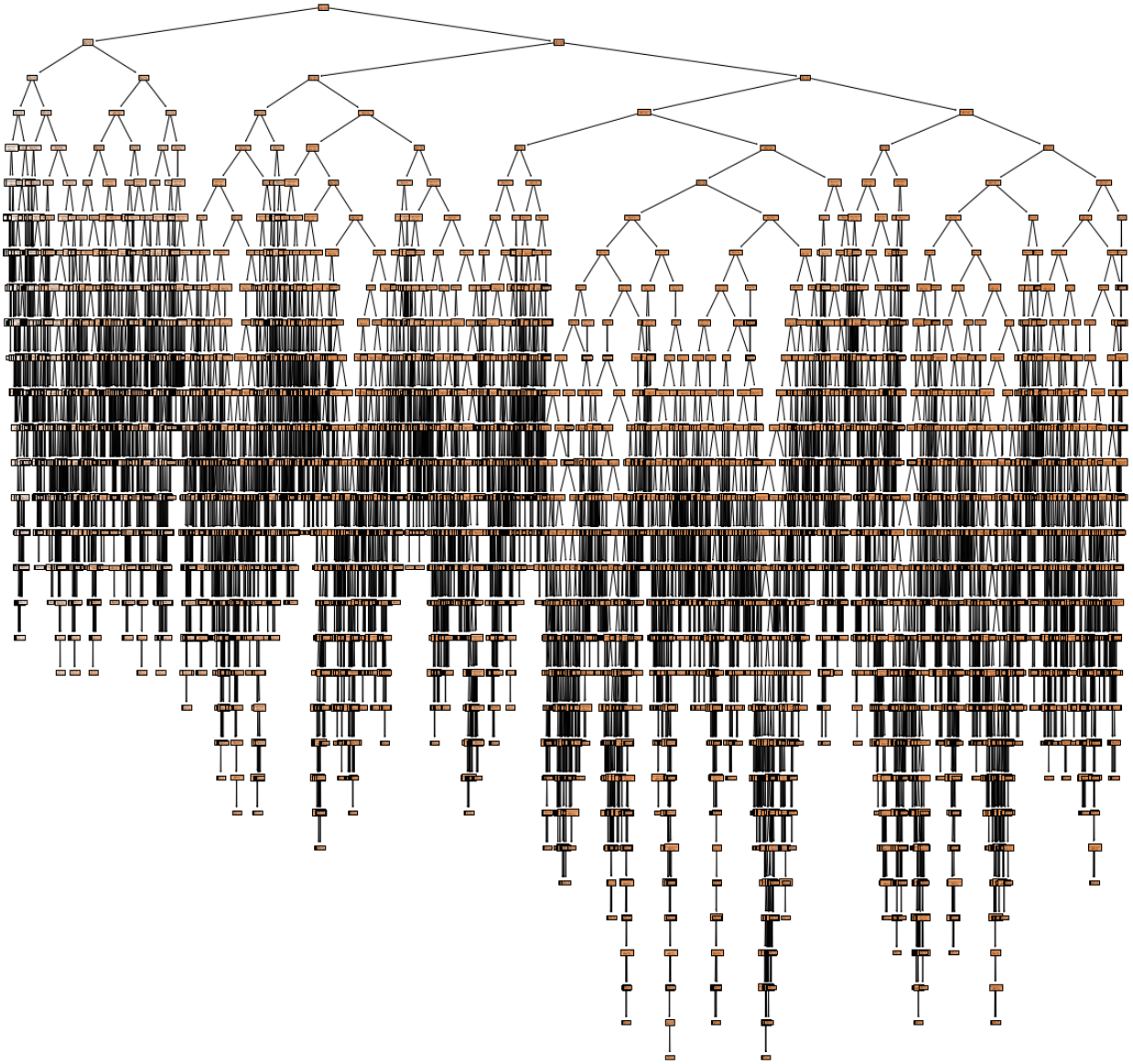


Boosting - reziduály

Regresory boli podobné výsledkami, najlepšie dosiahol **RandomForrest**, ktorý mal presnosť vyše 89%, tesne zaostával Boosting 88%. Reziduály v Boosting a RandomForest boli najviac vzdialené, jednalo sa len o minimálne počty (možno to pomohlo :D).

Dávid Gavenda 98533

BONUS – RandomForest – jeden strom



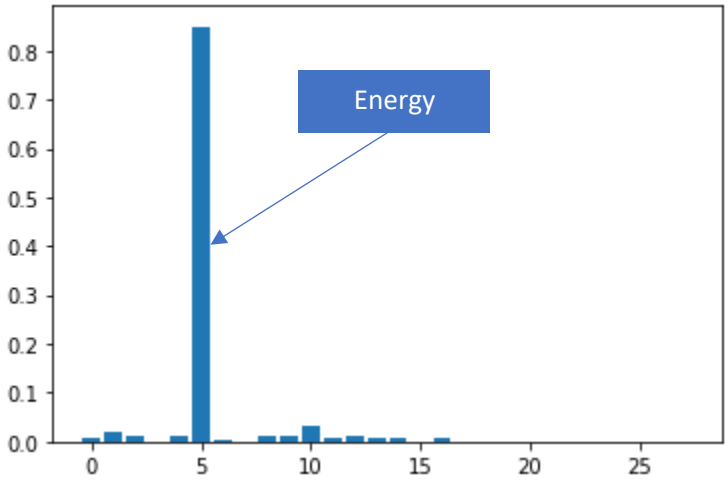
tree.png – jeden strom z RandomForest

Dávid Gavenda 98533

BONUS – RandomForest – sila vstupných príznakov

Feature: 0, Score: 0.00653	popularity	int64
Feature: 1, Score: 0.01966	release_date	int64
Feature: 2, Score: 0.01018	duration_ms	int64
Feature: 3, Score: 0.00021	explicit	int64
Feature: 4, Score: 0.01111	danceability	float64
Feature: 5, Score: 0.84939	**energy**	**float64**
Feature: 6, Score: 0.00459	key	int64
Feature: 7, Score: 0.00092	mode	int64
Feature: 8, Score: 0.01084	speechiness	float64
Feature: 9, Score: 0.01027	acousticness	float64
Feature: 10, Score: 0.03013	instrumentalness	float64
Feature: 11, Score: 0.00836	liveness	float64
Feature: 12, Score: 0.01018	valence	float64
Feature: 13, Score: 0.00845	tempo	float64
Feature: 14, Score: 0.00899	artist_followers	float64
Feature: 15, Score: 0.00035	alternative	int64
Feature: 16, Score: 0.00634	classical	int64
Feature: 17, Score: 0.00025	country	int64
Feature: 18, Score: 0.00000	drum and bass	int64
Feature: 19, Score: 0.00041	folk	int64
Feature: 20, Score: 0.00019	hip hop	int64
Feature: 21, Score: 0.00031	house	int64
Feature: 22, Score: 0.00013	metal	int64
Feature: 23, Score: 0.00101	pop	int64
Feature: 24, Score: 0.00019	rap	int64
Feature: 25, Score: 0.00057	rock	int64
Feature: 26, Score: 0.00042	soul	int64
Feature: 27, Score: 0.00000	trap	int64

Energy, ktorý nám vyšiel už z matice, má najvyššiu hodnotu (váhu), ostatné spomínané sú taktiež vyznačené. Medzi hodnoty, ktoré neboli predtým spomenuté ale mali vysokú váhu patria aj *release_date*, *duration_ms* a *speechiness*. Žánre celkovo moc vplyv nemali na hlasitosť (okrem *classical*).



Hodnoty z matice pre pripomenutie

danceability (0.42), energy(0.84), valence(0.37) a *acousticness(-0.72), instrumentalness(-0.51)* a *classical (-0.64)*

Dávid Gavenda 98533