Raport Analiza calității aerului și popularității știrilor online – EDA, Preprocesare și Compararea Algoritmilor de Clasificare

Titlu: „Analiza calității aerului și popularității știrilor online – EDA, Preprocesare și Compararea Algoritmilor de Clasificare”

Autori: Luca Plian

Data: 25 Mai 2025

Introducere

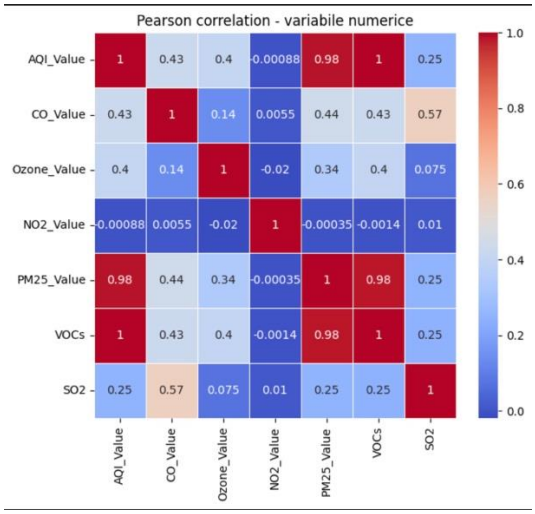
- În această lucrare investigăm două probleme de clasificare: estimarea nivelului calității aerului (AQI) și predicția popularității știrilor online. Vom parcurge etapele de EDA, preprocesare, antrenare și evaluare a patru algoritmi (Decision Tree, Random Forest, Logistic Regression, MLP).

1. Analiza dezechilibrului de clase

1.1 Identificarea tipurilor de atribute

- Numerice continue: AQI_Value, CO_Value, Ozone_Value, etc.
- Discrete: City, Country
- Ordinare: AQI_Category (Good < Moderate < … < Hazardous), Emissions(L0, L1, L2..), etc

Distincția dintre ordinale si discrete, consta in ca in ordinare sunt destul de putine valori unice, ceea ce ne permit in a contribui la clasificare, cele discrete le eliminam



Din heatmap se remarcă câteva relații foarte puternice (>= 0.9):

1. Corelația dintre AQI_Value și VOCs (1.0) și corelația dintre AQI_Value și PM25_Value (0.98): aceste perechi sunt practic redundante, indicând faptul că Compușii Organici Volatili (VOCs) și particule PM2.5 sunt factorii determinanți principali în calculul indexului general AQI.
2. Corelația dintre PM25_Value și VOCs (0.98): confirmă aceeași redundanță structurală între poluanți.

Există și corelații moderate spre puternice:

- Corelația dintre CO_Value și SO2 (0.57): un semnal de emisii chimice similare (probabil industriale sau trafic), dar suficient de slab pentru a păstra ambele în model, deoarece aduc nuanțe diferite.
- Corelația dintre AQI_Value și CO_Value (0.43) și corelația dintre AQI_Value și Ozone_Value (0.40): corelații semnificative dar nu copleșitoare, sugerând că acești poluanți contribuie la calitatea aerului, dar nu o definesc singuri.

Restul sunt aproape de zero, indicând o aparentă independență statistică față de indexul general în acest set de date

Astfel, heatmap-ul ne-a ghidat selecția finală a caracteristicilor înainte de antrenarea oricărui model.

Column	Number of non-nan values	Total number of values per column	Unique values
Country	18421	18770	176
City	18770	18770	18770
AQI_Value	18770	18770	327
CO_Value	18770	18770	34
CO_Category	16877	18770	3
Ozone_Value	16900	18770	212
Ozone_Category	18770	18770	5
NO2_Value	18770	18770	60
NO2_Category	18770	18770	2
PM25_Value	18770	18770	366
PM25_Category	18770	18770	6
VOCs	18770	18770	18770
SO2	18770	18770	18770
Emissions	18770	18770	6
AQI_Category	18770	18770	6

Analiza Variabilelor Lipsă (Missing Values) - Air

Majoritatea coloanelor au 18770 rânduri, ce este dimensiunea totală a setului de date.

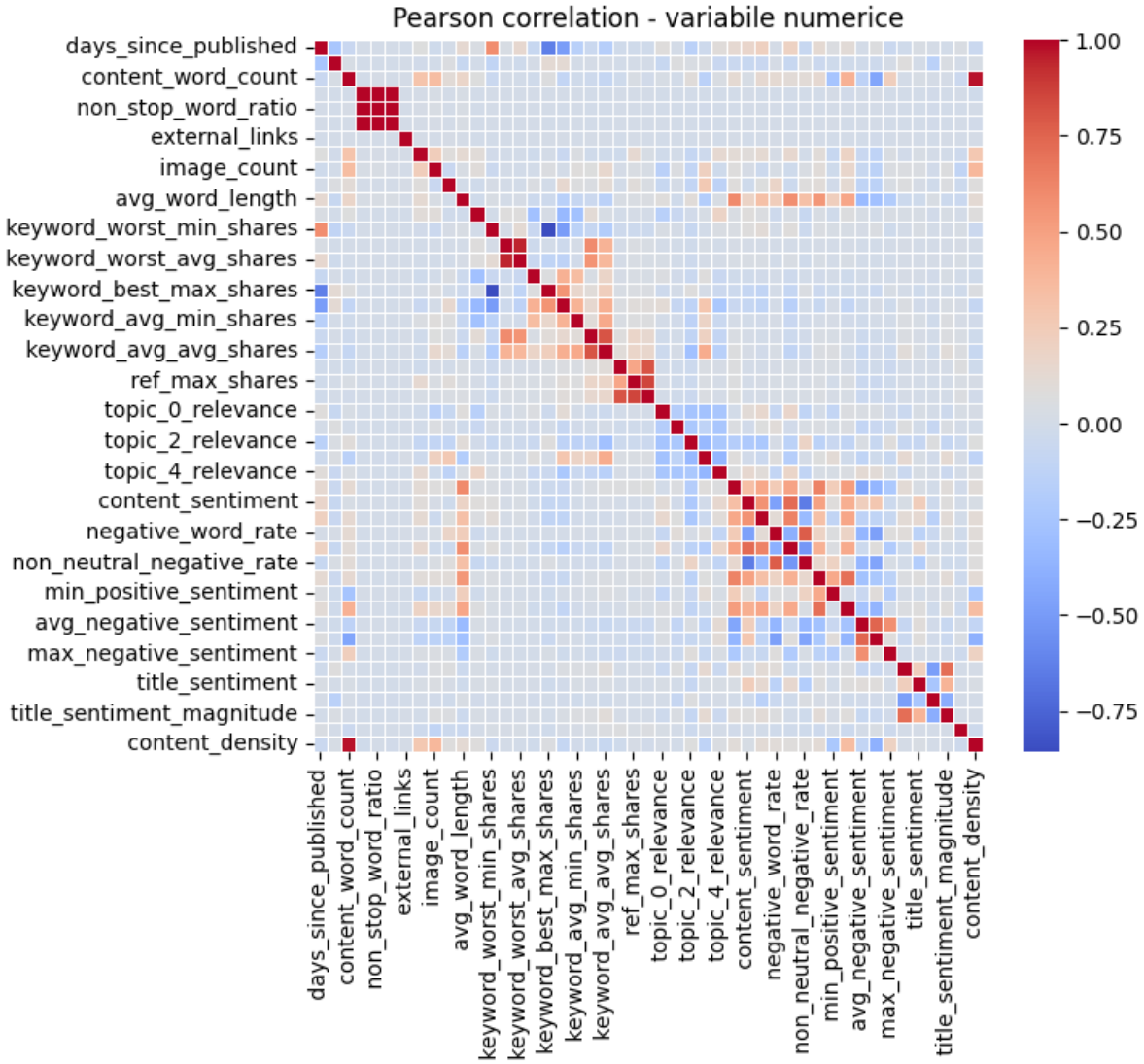
- Country: Are 18.421 valori non-nan, deci lipsesc 349 de valori.
- CO_Category: Are 16.877 valori, deci lipsesc 1.893 de înregistrări.
- Ozone_Value: Are 16.900 valori, deci lipsesc 1.870 de înregistrări.

Column	Number of non-nan values	Total number of values per column	Unique values
channel_lifestyle	28540	31715	3
content_density	28570	31715	28571

Analiza Variabilelor Lipsă (Missing Values) - News

Majoritatea coloanelor au 31715 rânduri, ce este dimensiunea totală a setului de date.

- Channel_lifestyle: Are valori 28540 non-nan, deci lipsesc 3175 de valori.
- content_density: Are 28570 valori, deci lipsesc 3145 de înregistrări.



Din heatmap se remarcă câteva relații foarte puternice (culori extreme, roșu închis sau albastru închis):

1. Grupul de metrici de complexitate a textului (Roșu Intens): Se observă un bloc compact de corelație pozitivă aproape perfectă între non_stop_word_ratio, unique_word_ratio și unique_non_stop_ratio. Aceste variabile sunt practic redundante, măsurând variații ale aceluiași concept (diversitatea vocabularului); dacă una crește, celelalte cresc automat.
2. Grupul statisticilor despre cuvinte cheie (Roșu Intens): Variabilele care încep cu kw_ (ex. kw_avg_avg, kw_max_avg) formează un alt cluster de corelație pozitivă puternică. Articolele care conțin cuvinte cheie populare tind să aibă valori ridicate la toate metricile asociate (min, max, avg shares).
3. Relația inversă densitate-lungime (Albastru Intens): Se distinge o corelație negativă puternică între content_word_count și content_density. Cu cât un articol este mai lung (mai multe cuvinte), cu atât densitatea informațională calculată tinde să scadă matematic.

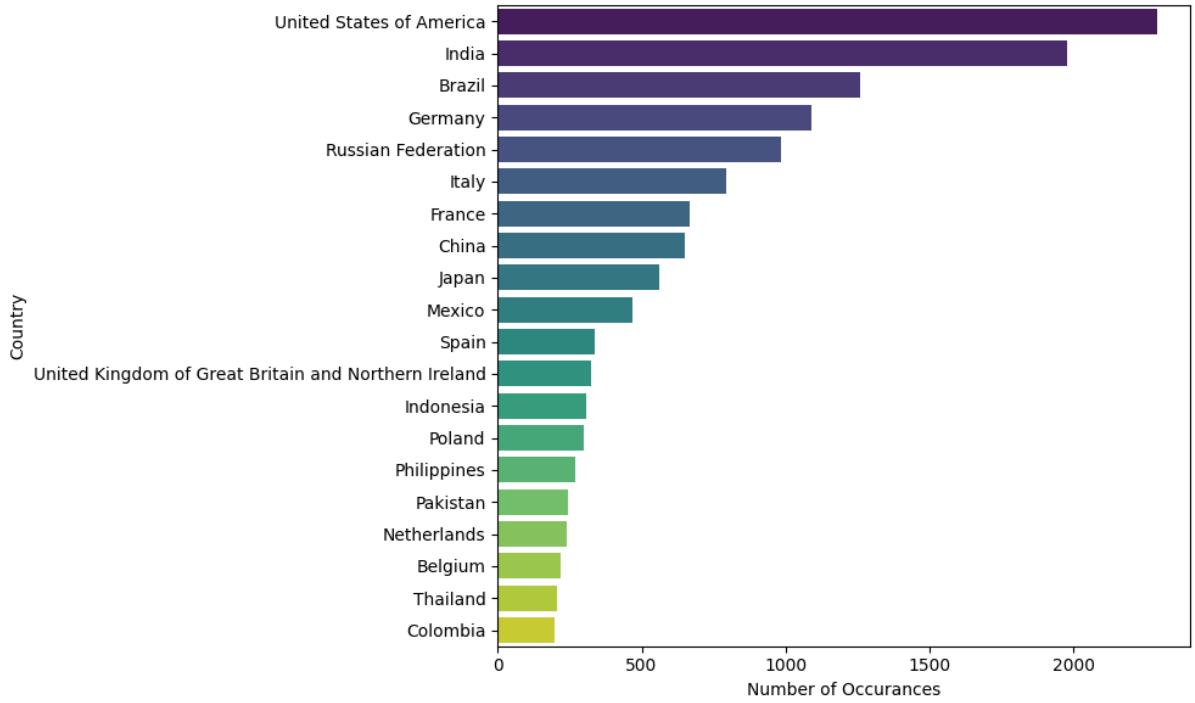
Barplots

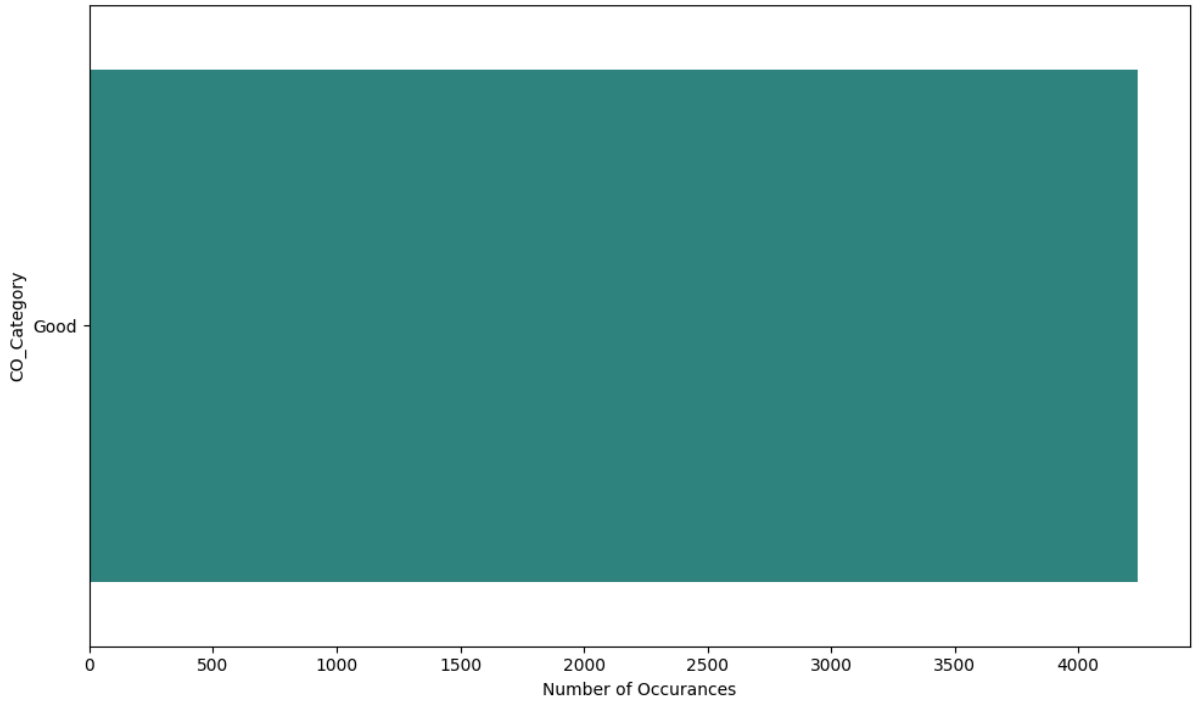
Analiza vizuală integrată a diagramelor de bare (Barplots) pentru variabilele țintă (AQI_Category) și cele explicative (Emissions, PM25_Category, Country) relevă o structură comună fundamentală: o distribuție de tip "Long-Tail" (Coadă Lungă), caracterizată prin dezechilibre extreme.

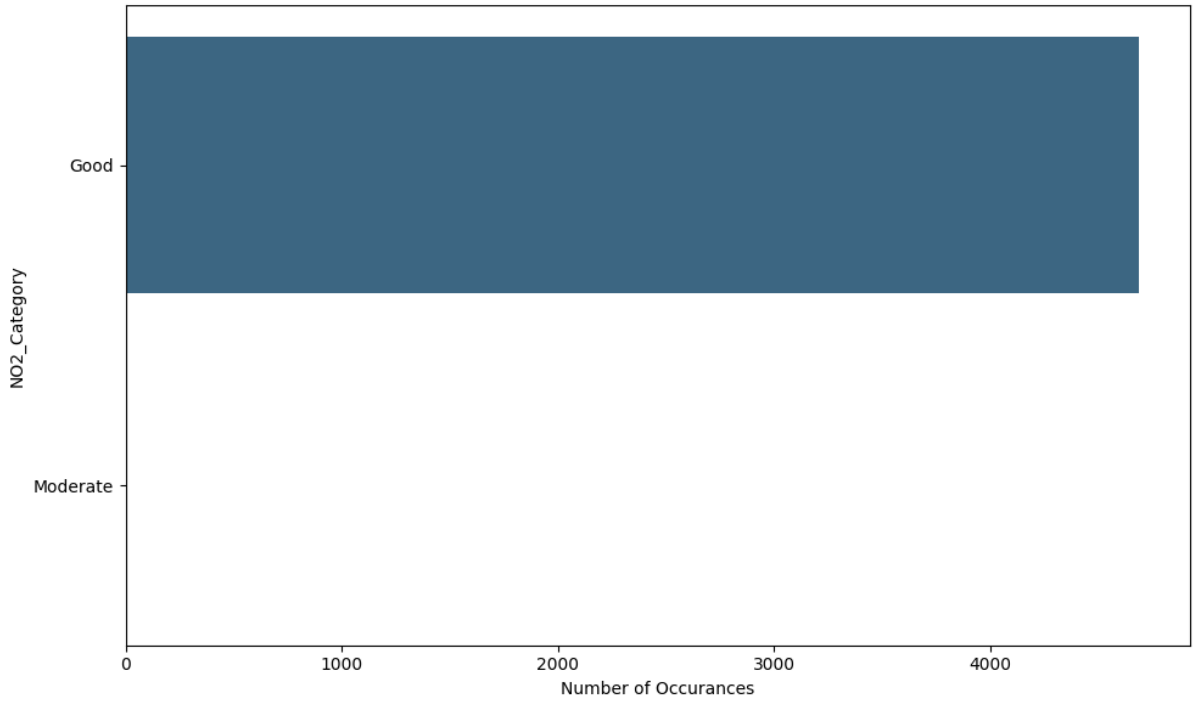
Pentru majoritatea variabilelor ordinale (PM25_Category, Emissions, Ozone_Category, AQI_Category), clasele asociate siguranței ("Good", "Moderate", "L0", "L1") domină vizual, cumulând majoritatea observațiilor. În contrast, stările critice ("Hazardous", "L5") sunt aproape imperceptibile grafic. Pentru NO2_Category, “Good” domină, iar „Moderate” este aproape inexistent, restul nu există. Pentru CO_Category, singura categorie este “Good”. Această structură confirmă consistența fizică a datelor: episoadele de poluare severă sunt evenimente rare, constituind excepția și nu regula.

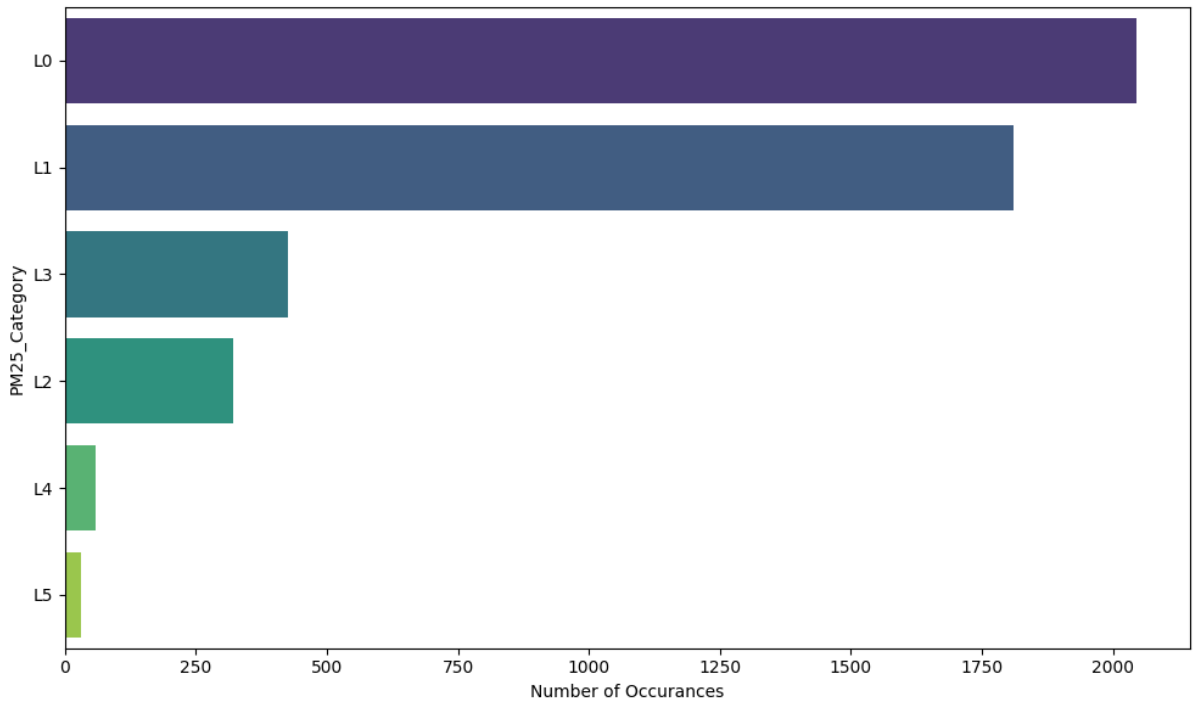
Pentru Country, graficul este dominat de doar doi actori majori (SUA și India), restul țărilor formând o "coadă lungă" de categorii slab reprezentate. Acest lucru indică faptul că setul de date nu este uniform distribuit global, ci este puternic dominat de anumite țări.

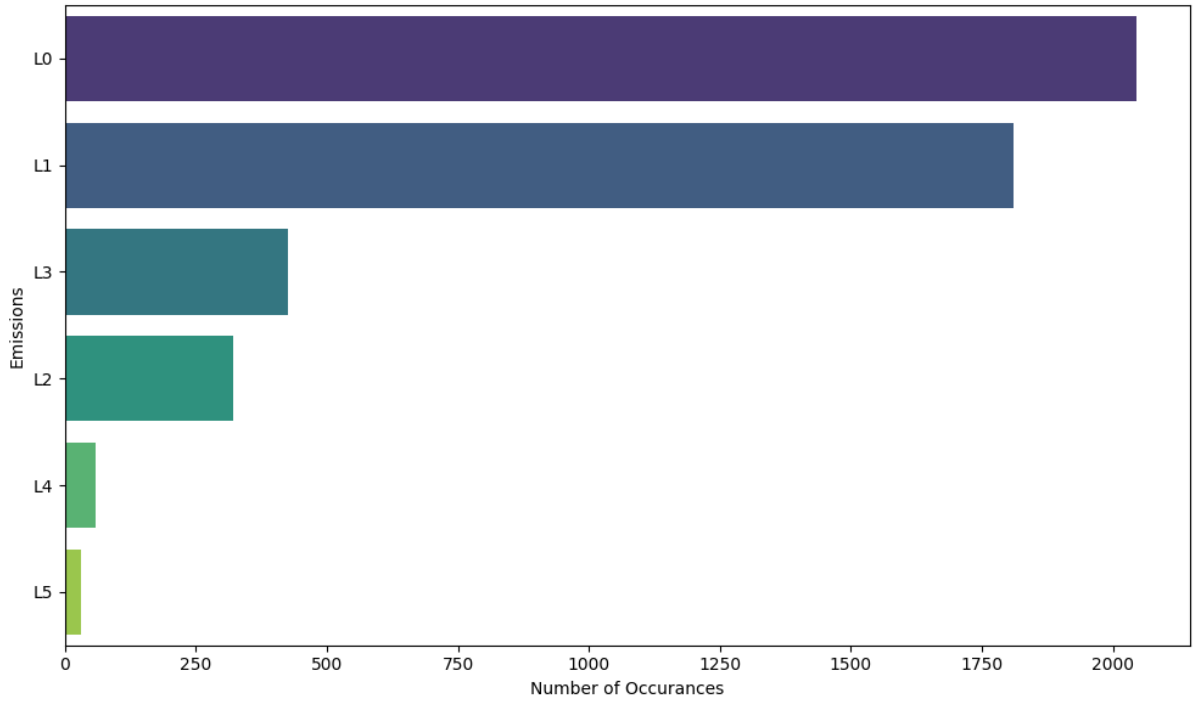
Dominanța vizuală a claselor majoritare sugerează un risc major de **bias (părtinire)**. Un model neponderat va tinde să optimizeze acuratețea globală ignorând complet clasele rare (poluarea extremă sau țările mici), considerându-le zgomot statistic. Această analiză vizuală justifică imperativ necesitatea strategiilor de re-eșantionare și utilizarea parametrului class_weight='balanced' pentru a forța algoritmii să acorde atenție evenimentelor critice.

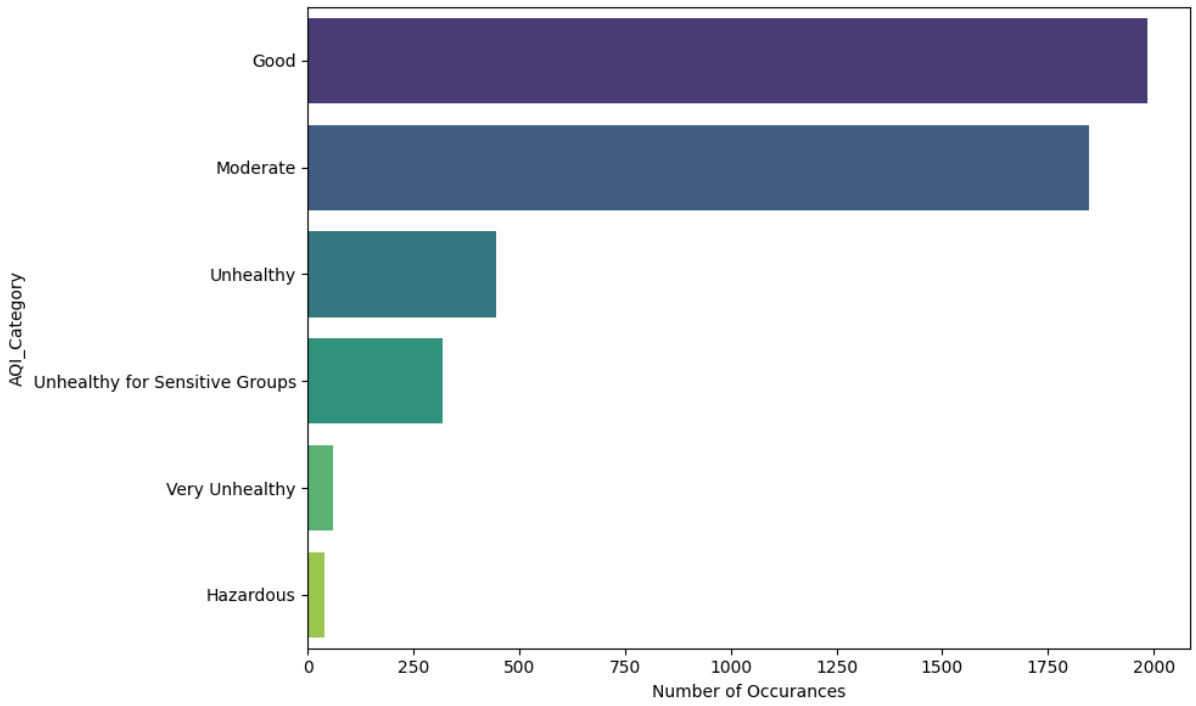


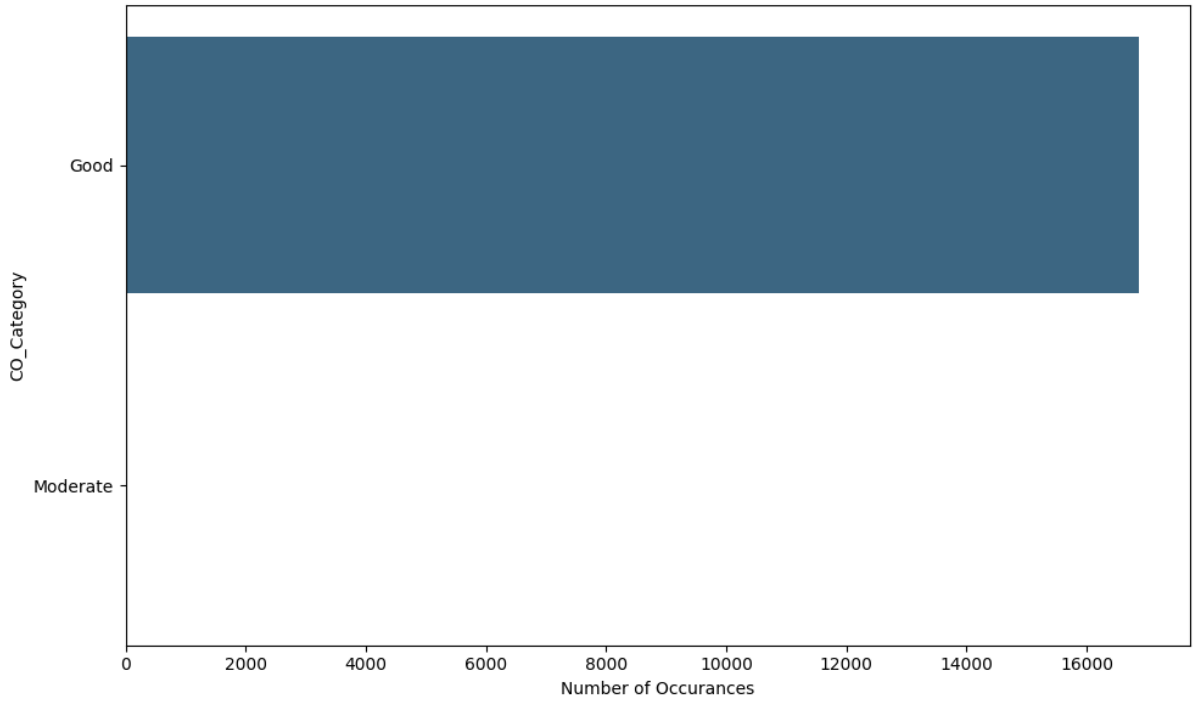

CO_Category
Good
0
500
1000
1500
2000
2500
3000
3500
4000
Number of Occurances

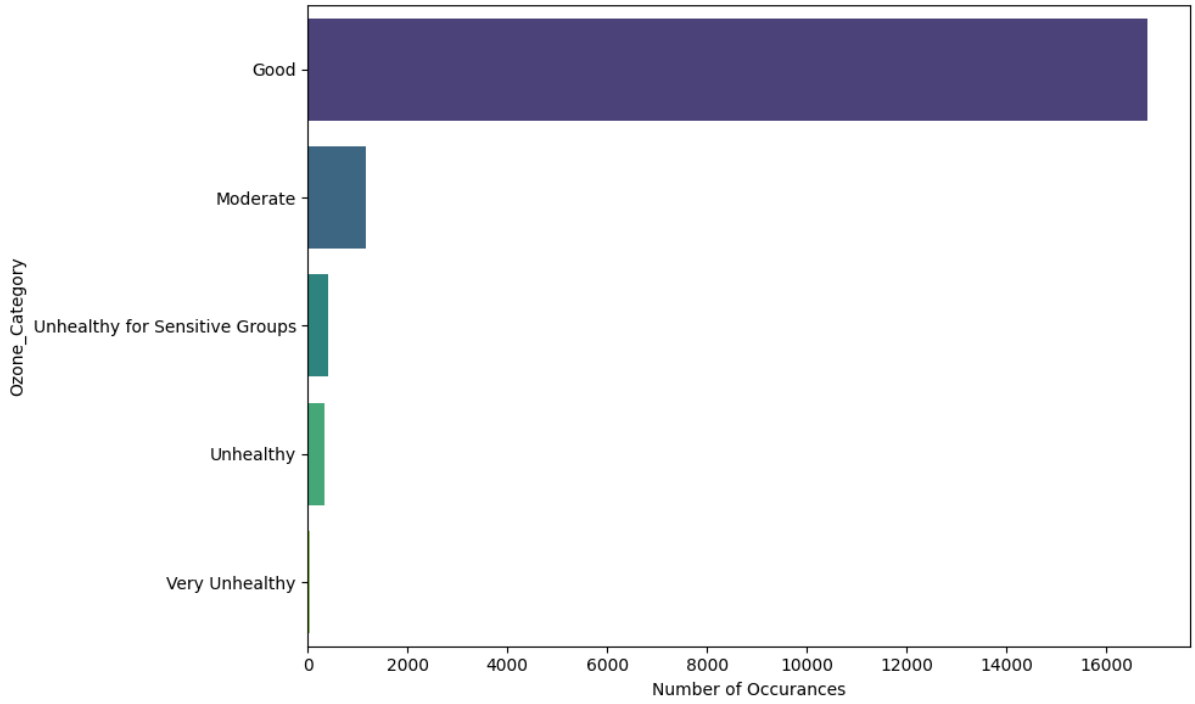

NO2_Category
Good
Moderate
0
1000
2000
3000
4000
Number of Occurances


L0
L1
L3
L2
L4
L5
PM25_Category
0
250
500
750
1000
1250
1500
1750
2000
Number of Occurances


L0
L1
L3
L2
L4
L5
Emissions
0
250
500
750
1000
1250
1500
1750
2000
Number of Occurances

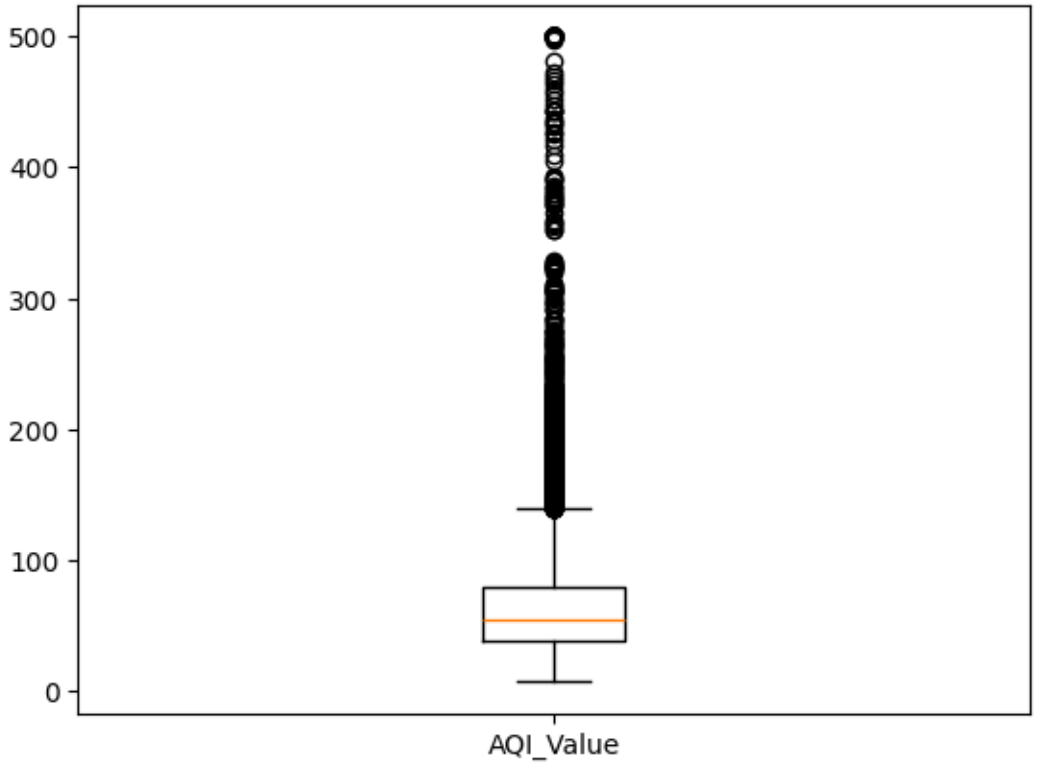

Good
Moderate
Unhealthy
Unhealthy for Sensitive Groups
Very Unhealthy
Hazardous
AQI_Category
0
250
500
750
1000
1250
1500
1750
2000
Number of Occurances

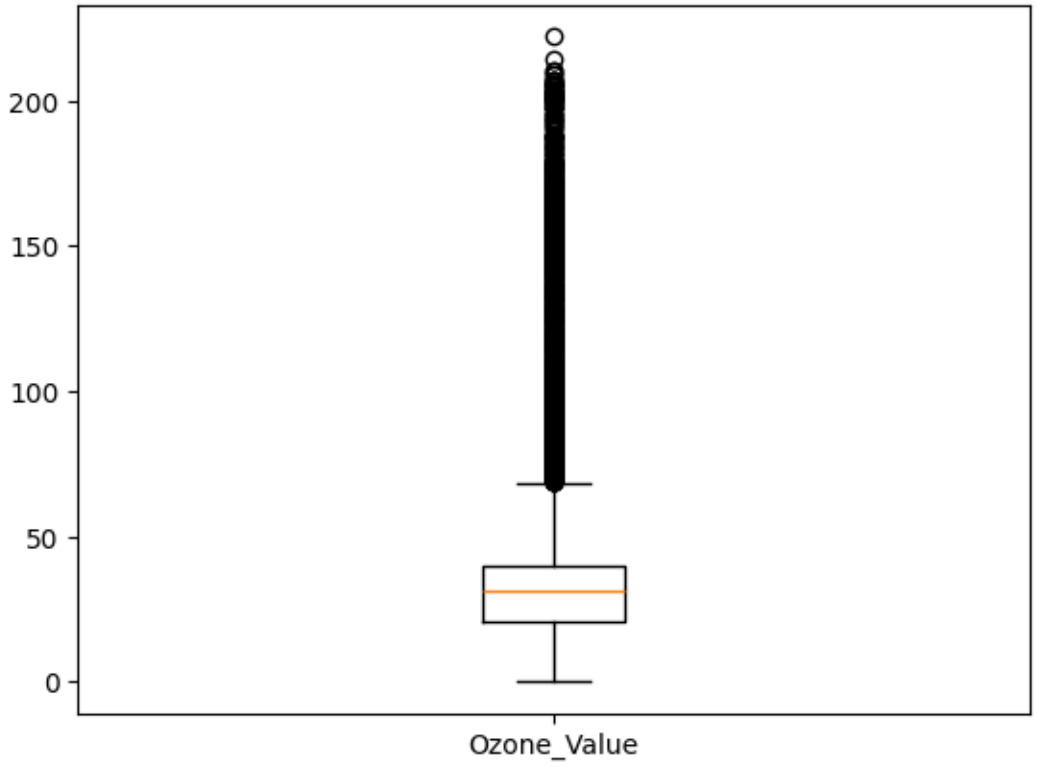

Good
Moderate
CO_Category
0
2000
4000
6000
8000
10000
12000
14000
16000
Number of Occurances

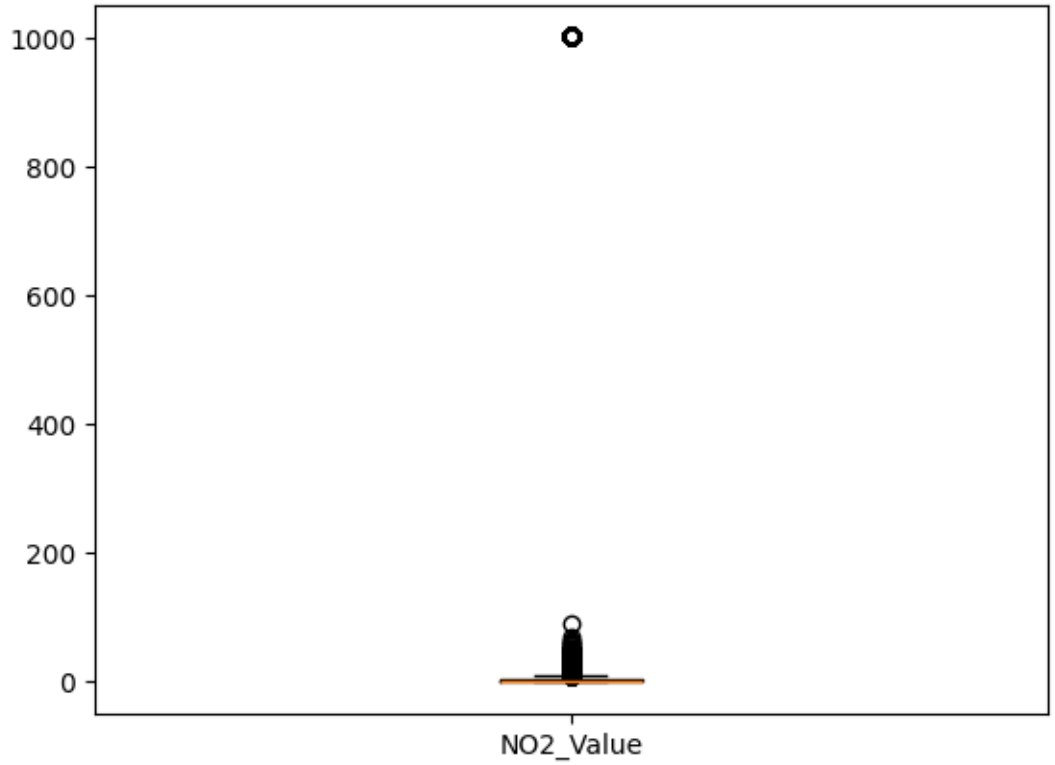


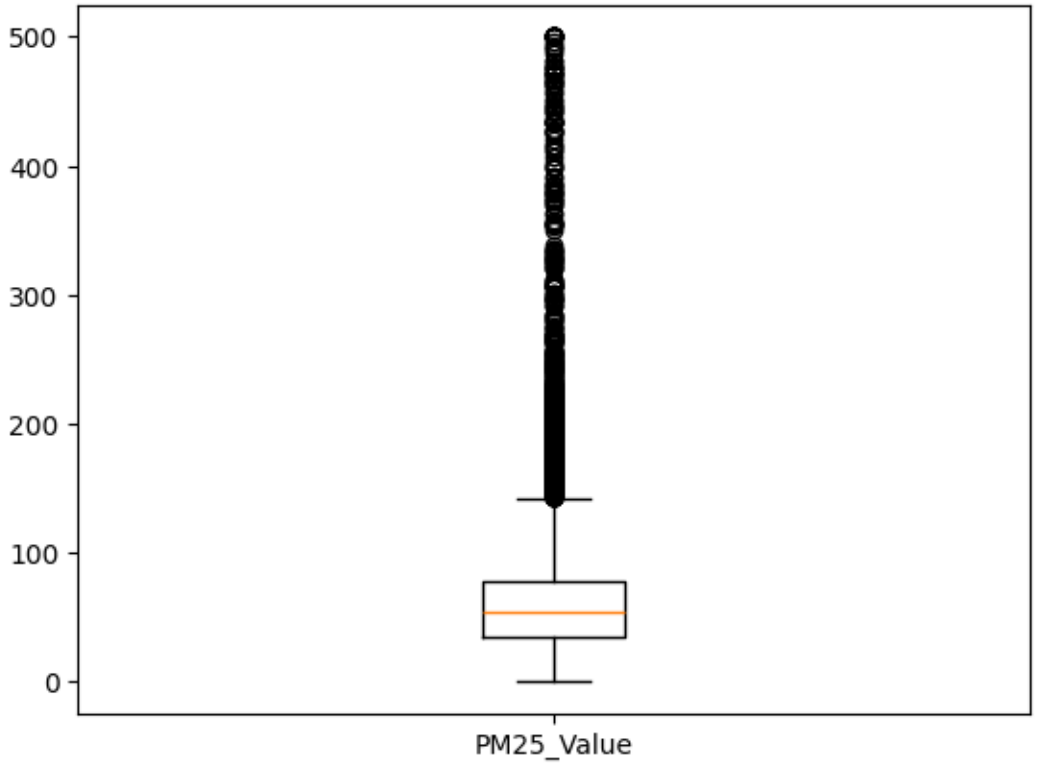
Boxplot

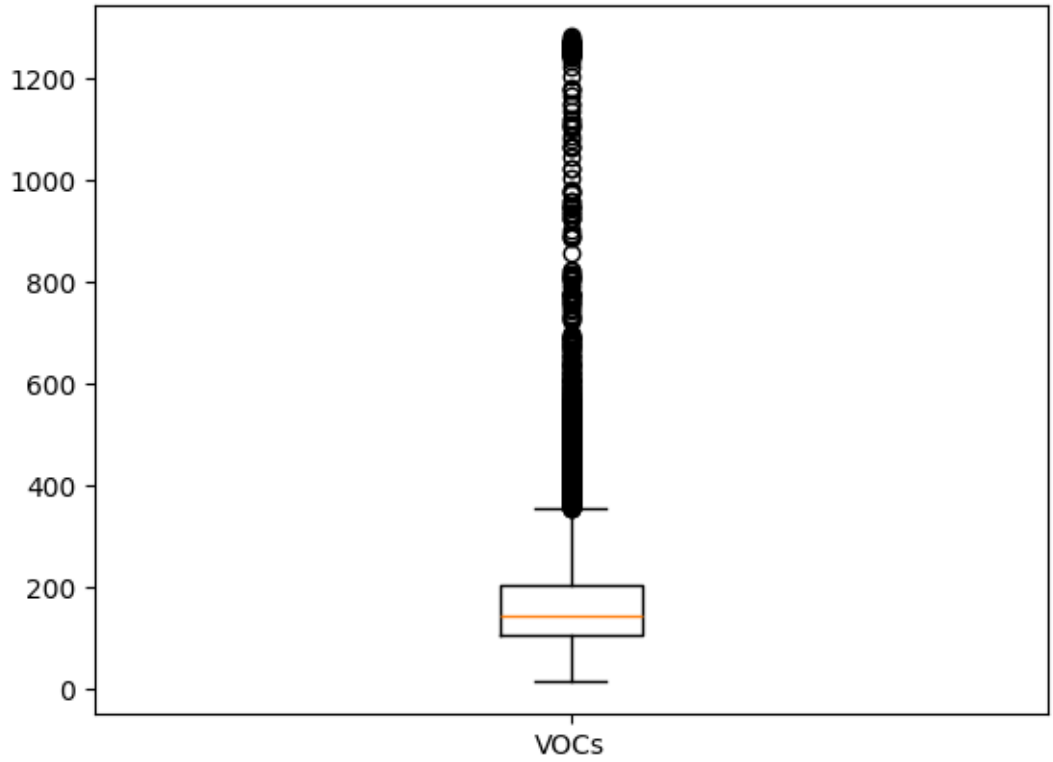
Examinând Boxplot-urile generate pentru poluanții (în special CO_Value, PM2.5, NO2), observăm următoarele caracteristici structurale:
Cutia (IQR) Comprimată: Pentru majoritatea poluanților, box-ul (care reprezintă intervalul interquartilic, între Q1 și Q3) este foarte îngustă și situată în partea de jos a graficului. Acest lucru indică faptul că 50% din date sunt concentrate în valori mici.
Mustățile (Whiskers) și Valorile Extreme:
Se observă o mulțime densă de puncte situate mult deasupra "mustății" superioare (Upper Whisker).
În cazul particulelor PM2.5, punctele de outlier se extind foarte mult, indicând prezența unor valori extreme care sunt de câteva ori mai mari decât media.
Aceste puncte izolate confirmă prezența outlierilor semnificativi, care distorsionează media și deviația standard.
Concluzie vizuală: Distribuția datelor este puternic asimetrică (skewed-right). Prezența masivă a punctelor negre (outlieri) deasupra limitei superioare justifică

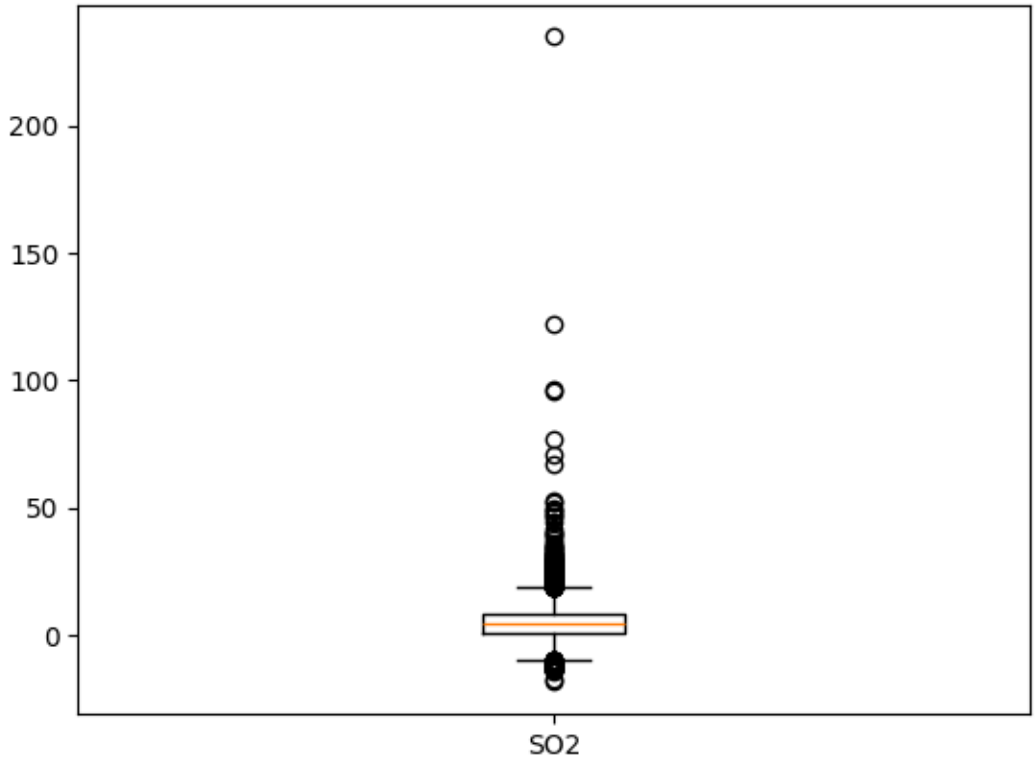

500
400
300
200
100
0
AQI_Value


200
150
100
50
0
Ozone_Value

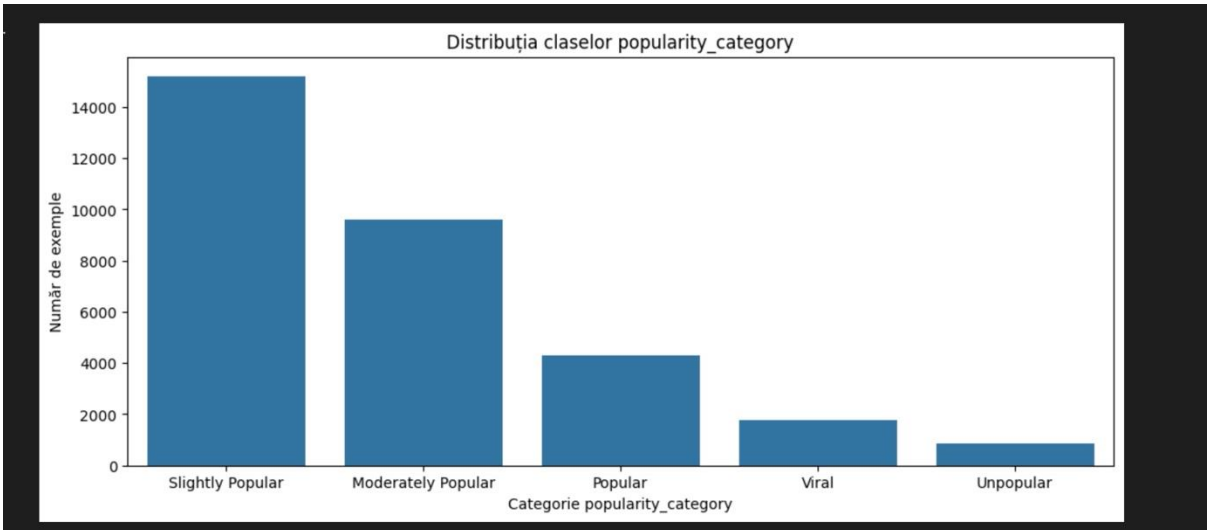

1000
800
600
400
200
0
NO2_Value


500
400
300
200
100
0
PM25_Value





Pentru News



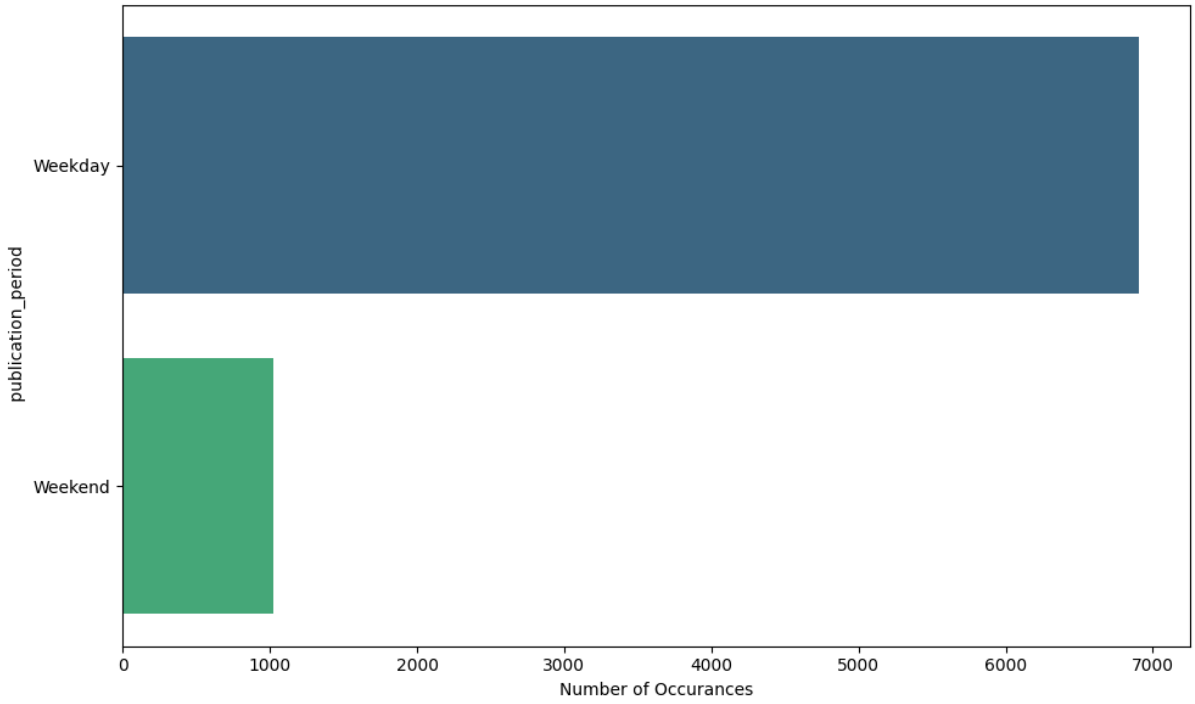
	Clasă	Frecvență	Procentaj (%)
0	Slightly Popular	15194	47.907930
1	Moderately Popular	9605	30.285354
2	Popular	4297	13.548794
3	Viral	1748	5.511588
4	Unpopular	871	2.746335

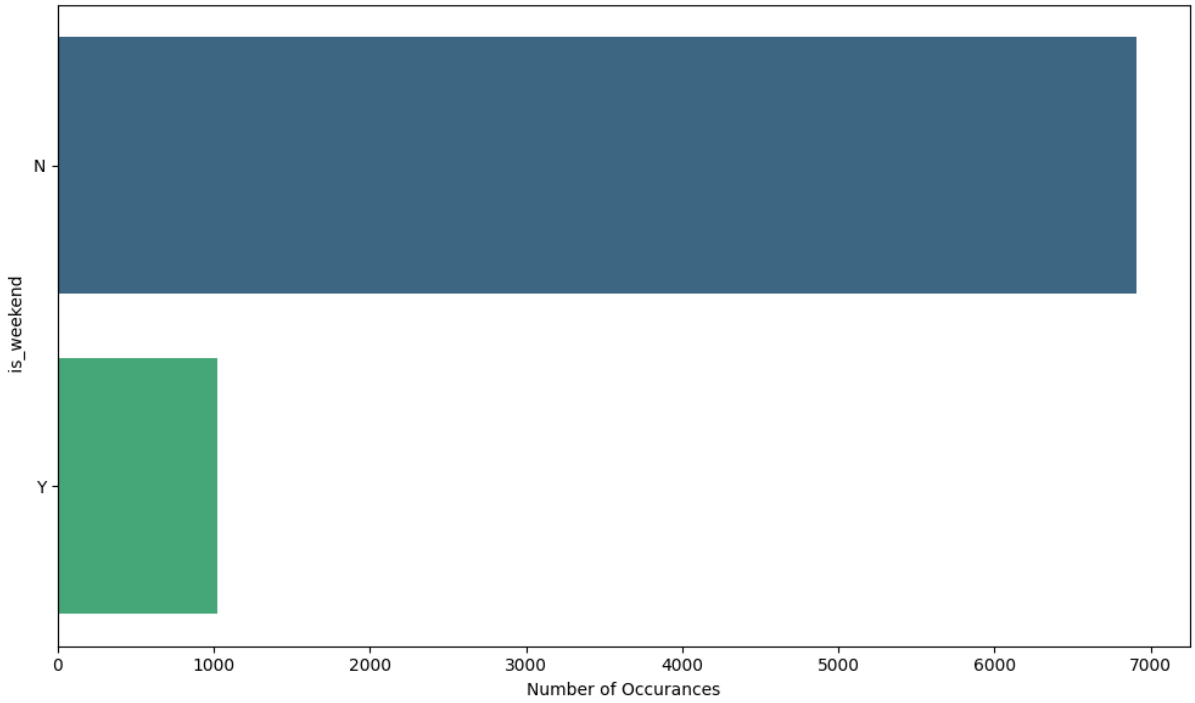
Analiza vizuală integrată a diagramelor de bare (Barplots) pentru variabila țintă (popularity_category) și celelalte atribute (Canalele de știri, Zilele săptămânii) relevă o structură comună fundamentală: o distribuție de tip "Long-Tail" (Coadă Lungă), caracterizată prin dezechilibre specifice fenomenelor de viralitate online.

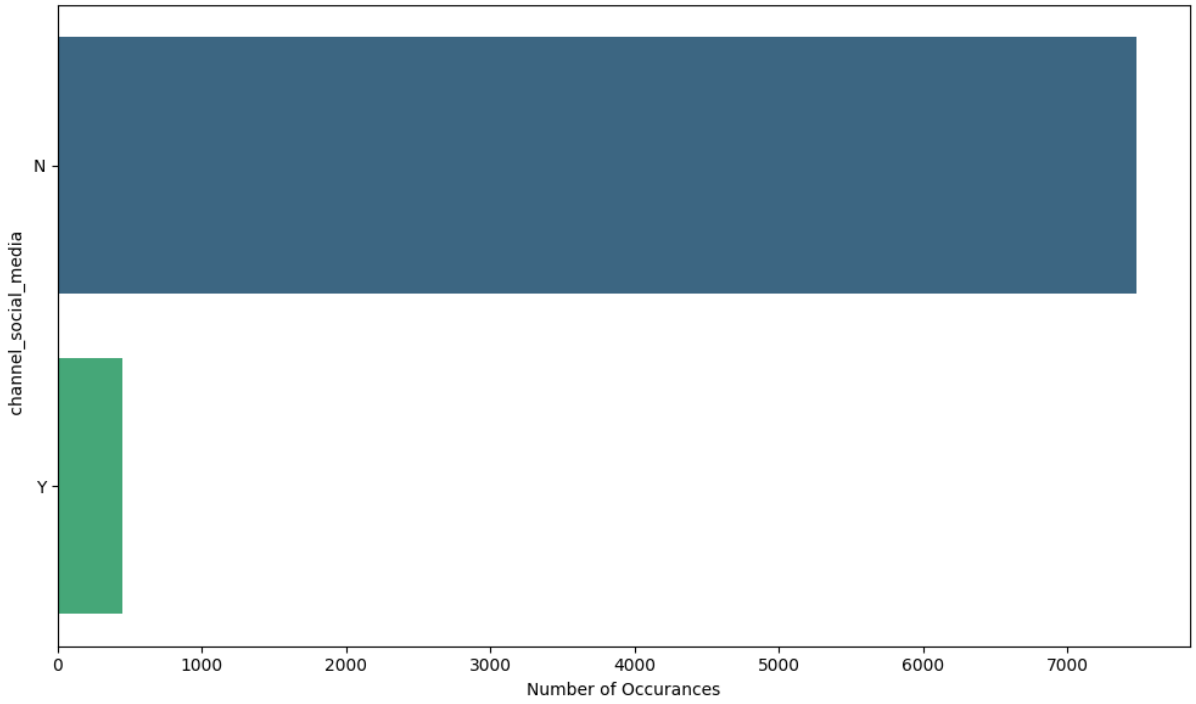
Pentru variabila țintă popularity_category, clasele de volum mediu ("Slightly Popular", "Moderately Popular") domină vizual, cumulând majoritatea observațiilor. În contrast, stările extreme ("Unpopular", "Viral") sunt mult mai puțin vizibile grafic, apărând ca bare reduse. Această structură confirmă natura stocastică a știrilor: succesul masiv (Viral) sau eșecul total sunt evenimente rare, constituind excepția și nu regula.

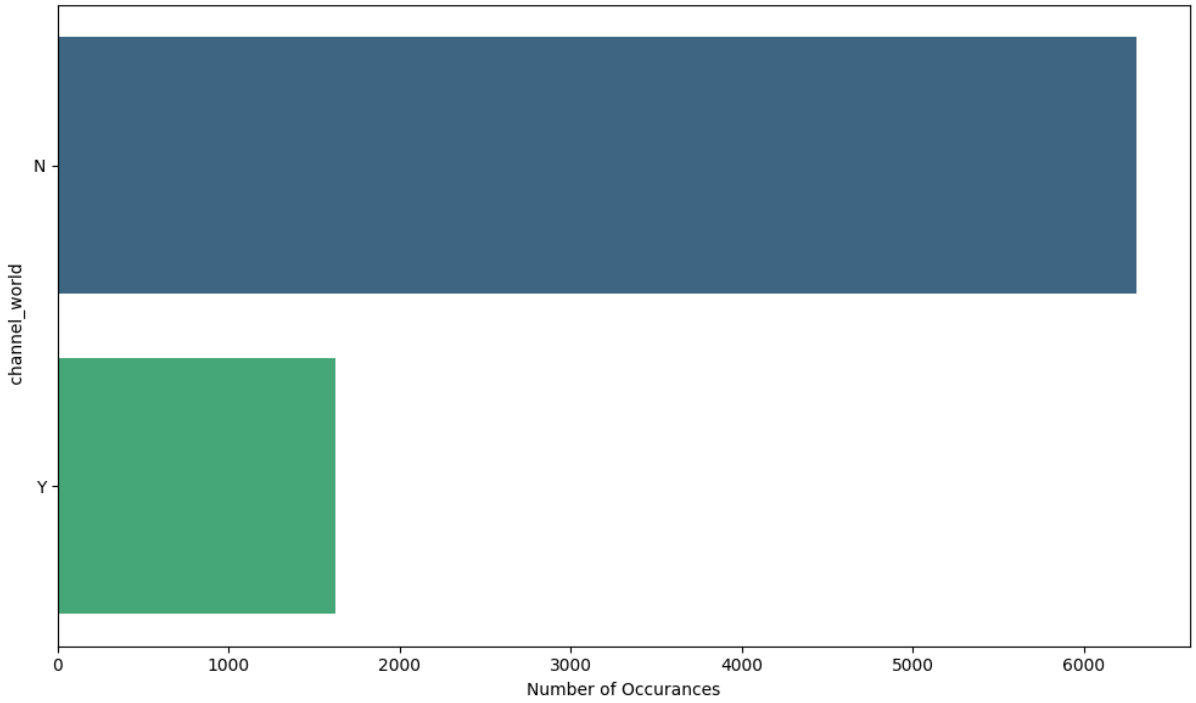
Pentru atributele categorice (Canale și Timp), graficele sunt dominate de anumite categorii editoriale. Canalele "World", "Tech" și "Entertainment" sunt actori majori, în timp ce canalele precum "Social Media" sau "Lifestyle" formează o categorie slab reprezentată. Similar, zilele lucrătoare (Weekday) domină vizual față de Weekend (care apare ca o categorie minoritară), indicând faptul că setul de date nu este uniform distribuit în timp, ci reflectă ciclul de lucru standard al redacțiilor.

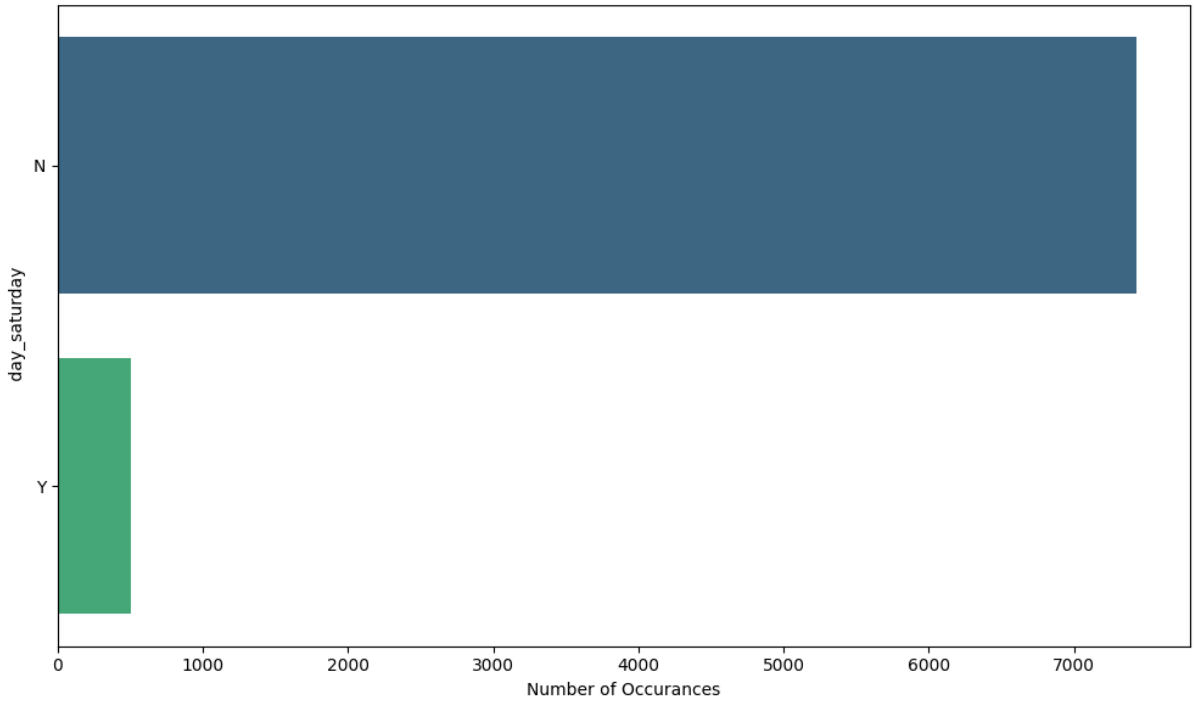
Dominanța vizuală a claselor majoritare sugerează un risc major de bias (părtinire). Un model neponderat va tinde să optimizeze acuratețea globală ignorând complet clasele rare (articolele virale sau cele publicate în weekend), considerându-le zgomot statistic. Această analiză vizuală justifică imperativ necesitatea strategiilor de re-eșantionare și utilizarea parametrului class_weight='balanced' pentru a forța algoritmii să acorde atenție articolelor cu potențial ridicat.

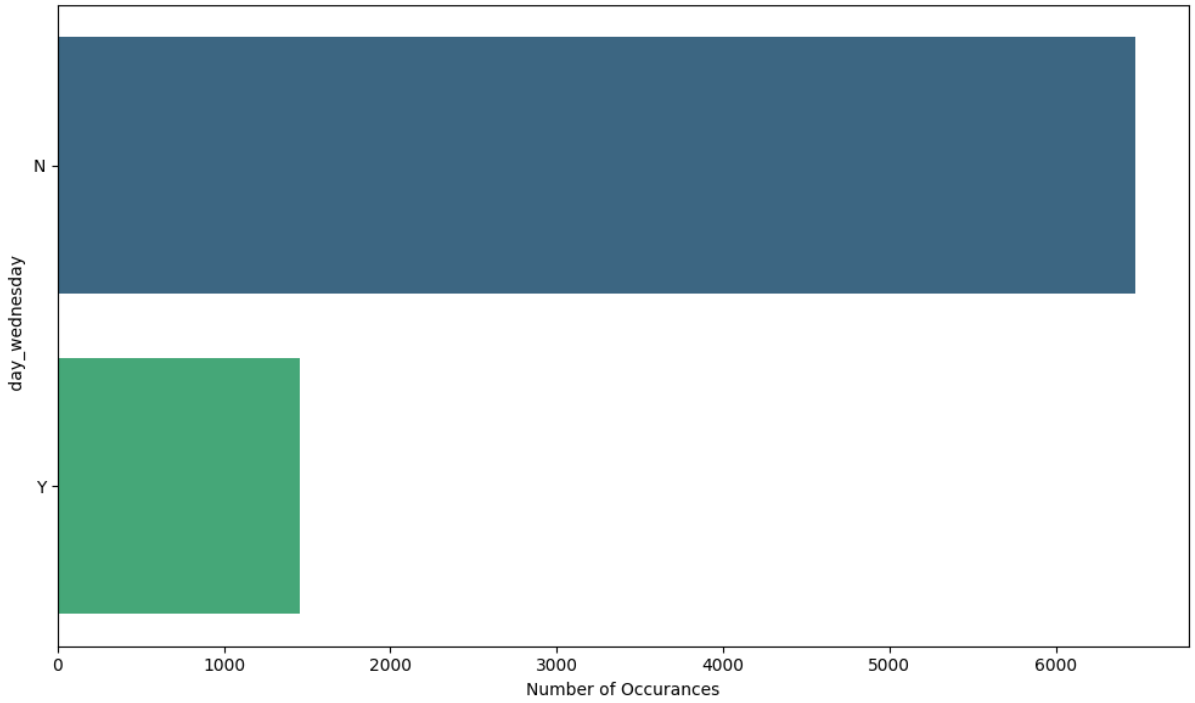

Weekday
Weekend
publication_period
0
1000
2000
3000
4000
5000
6000
7000
Number of Occurances

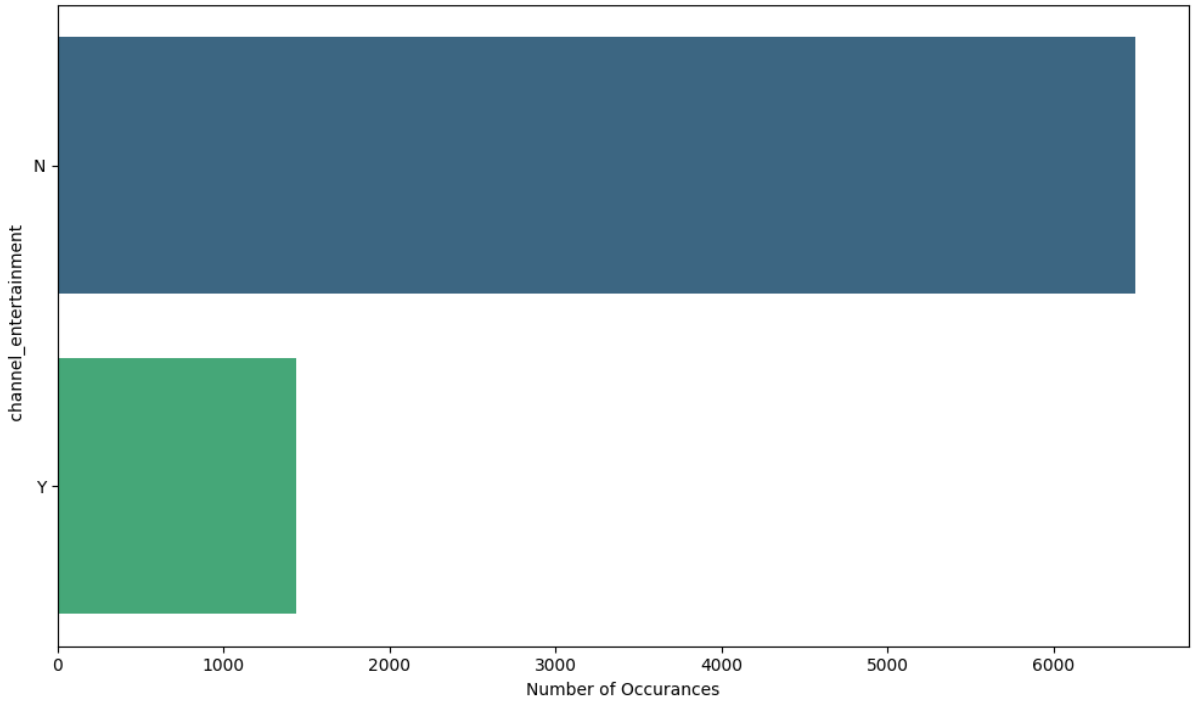

N
Y
is_weekend
0
1000
2000
3000
4000
5000
6000
7000
Number of Occurances

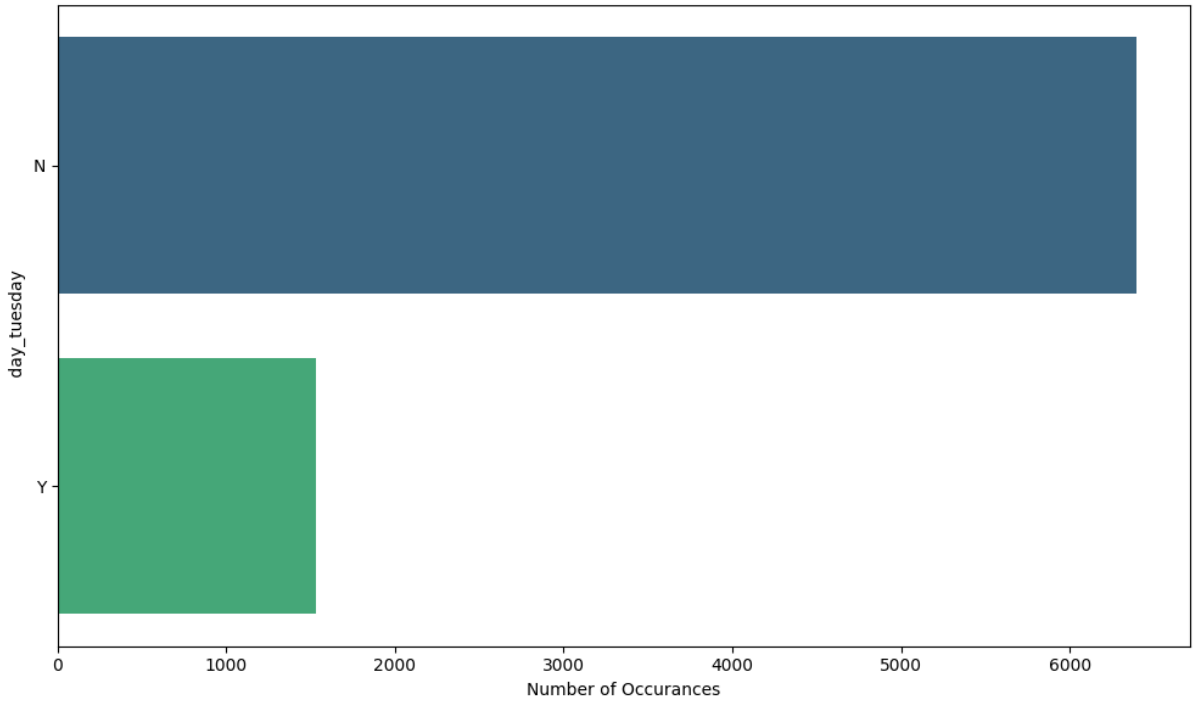

channel_social_media
N
Y
0
1000
2000
3000
4000
5000
6000
7000
Number of Occurances

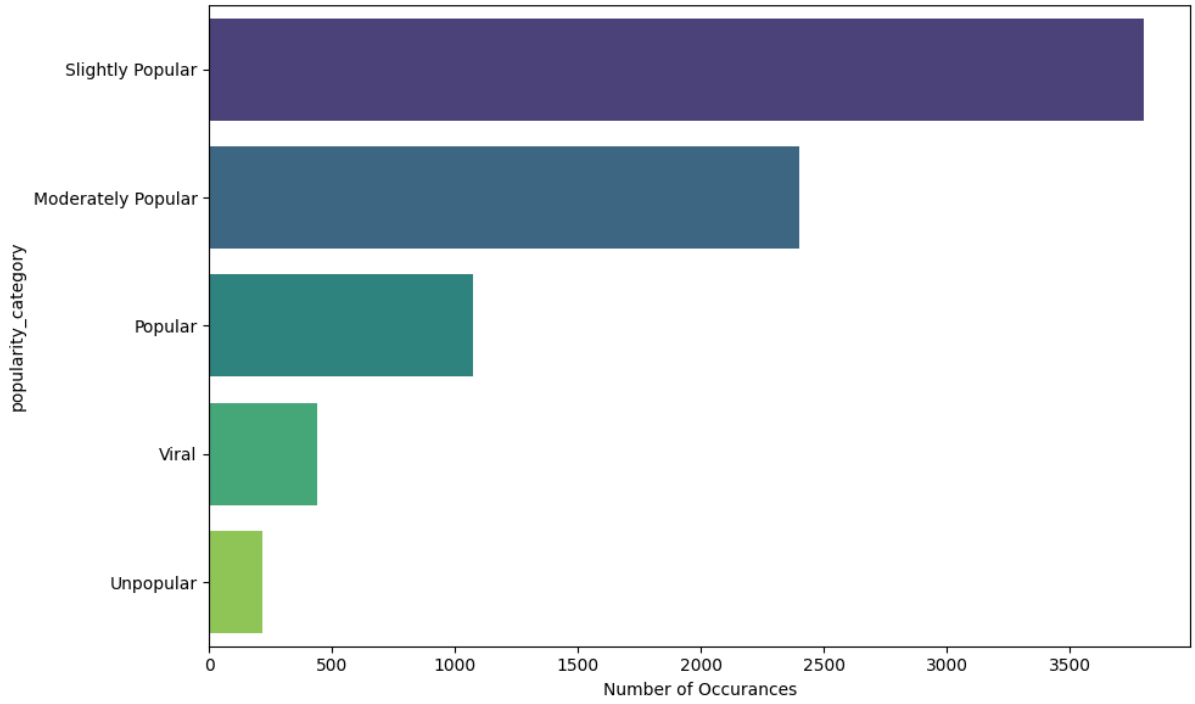

channel_world
N
Y
0
1000
2000
3000
4000
5000
6000
Number of Occurances

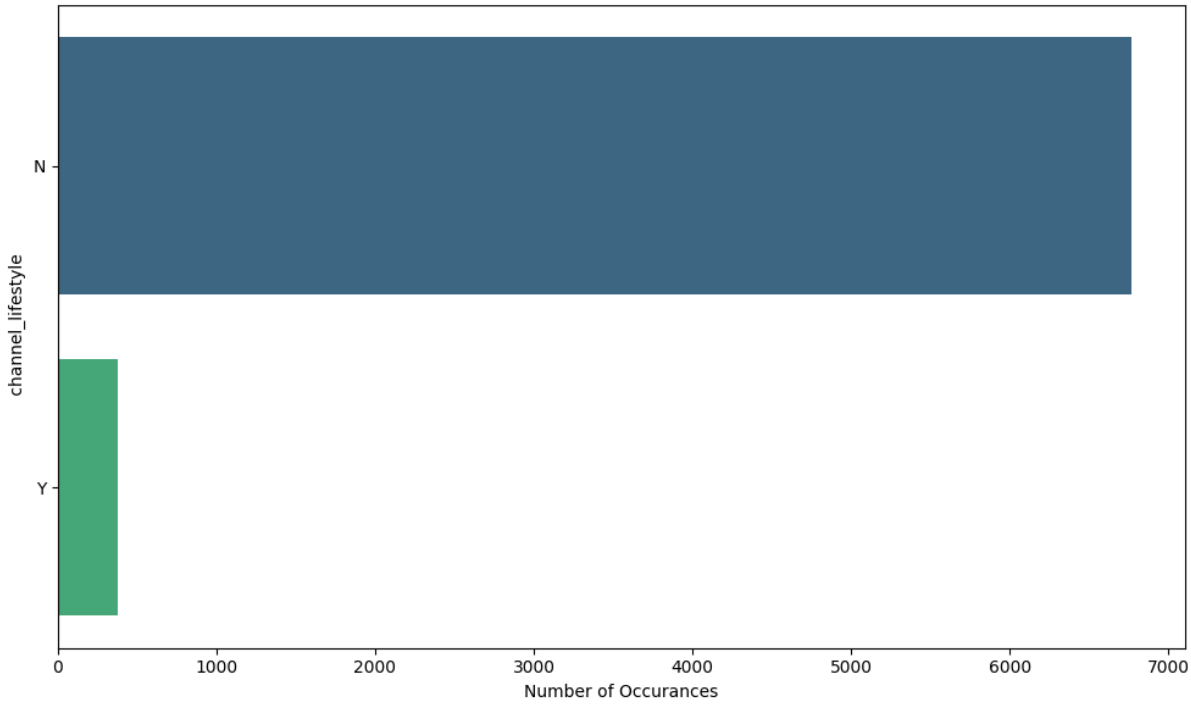

N
Y
day_saturday
0
1000
2000
3000
4000
5000
6000
7000
Number of Occurances

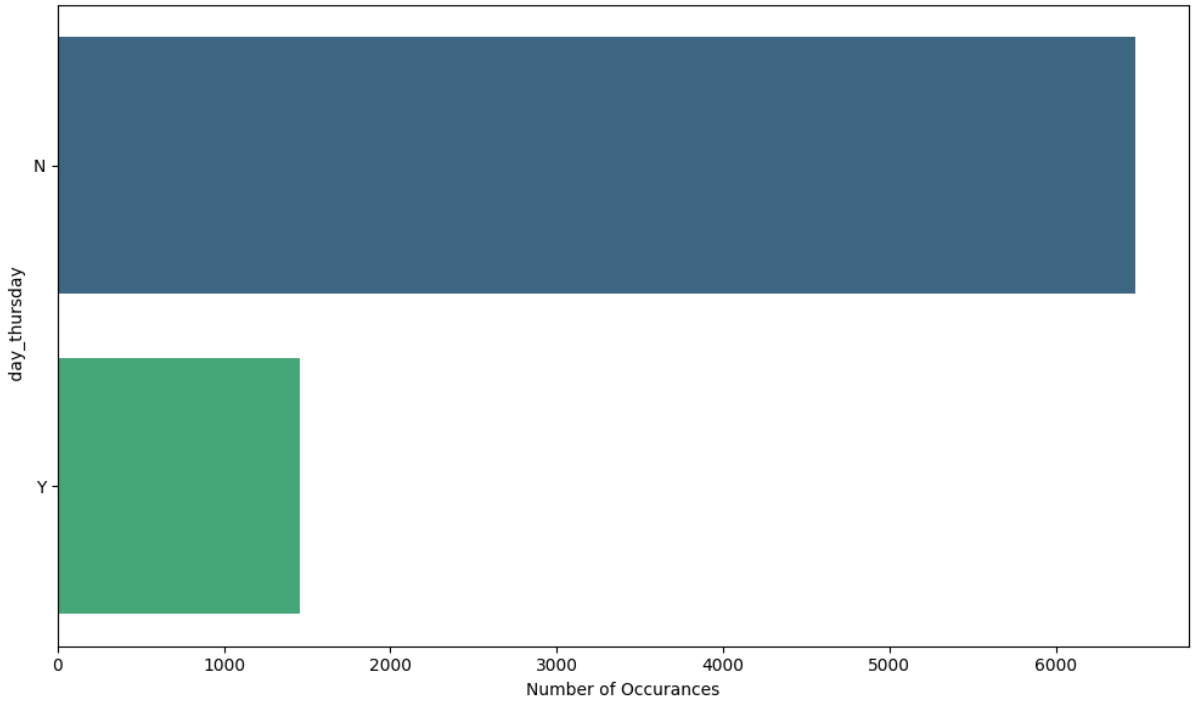

N
Y
day_wednesday
0
1000
2000
3000
4000
5000
6000
Number of Occurances

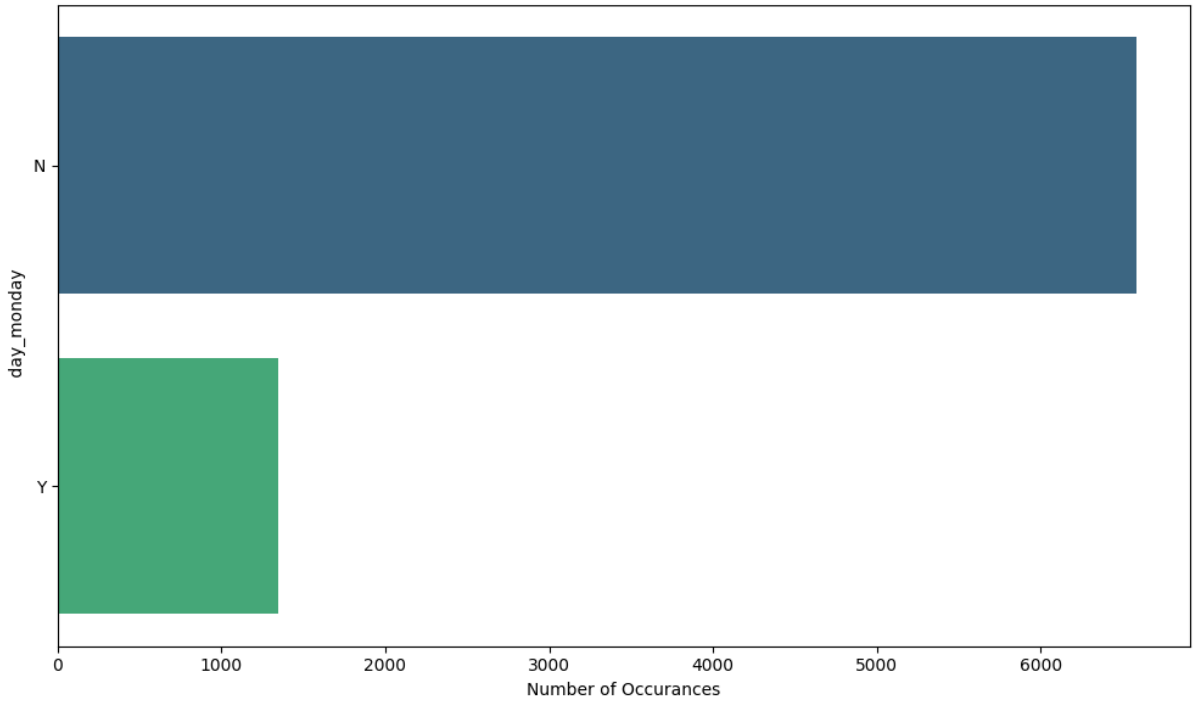

channel_entertainment
N
Y
0
1000
2000
3000
4000
5000
6000
Number of Occurances

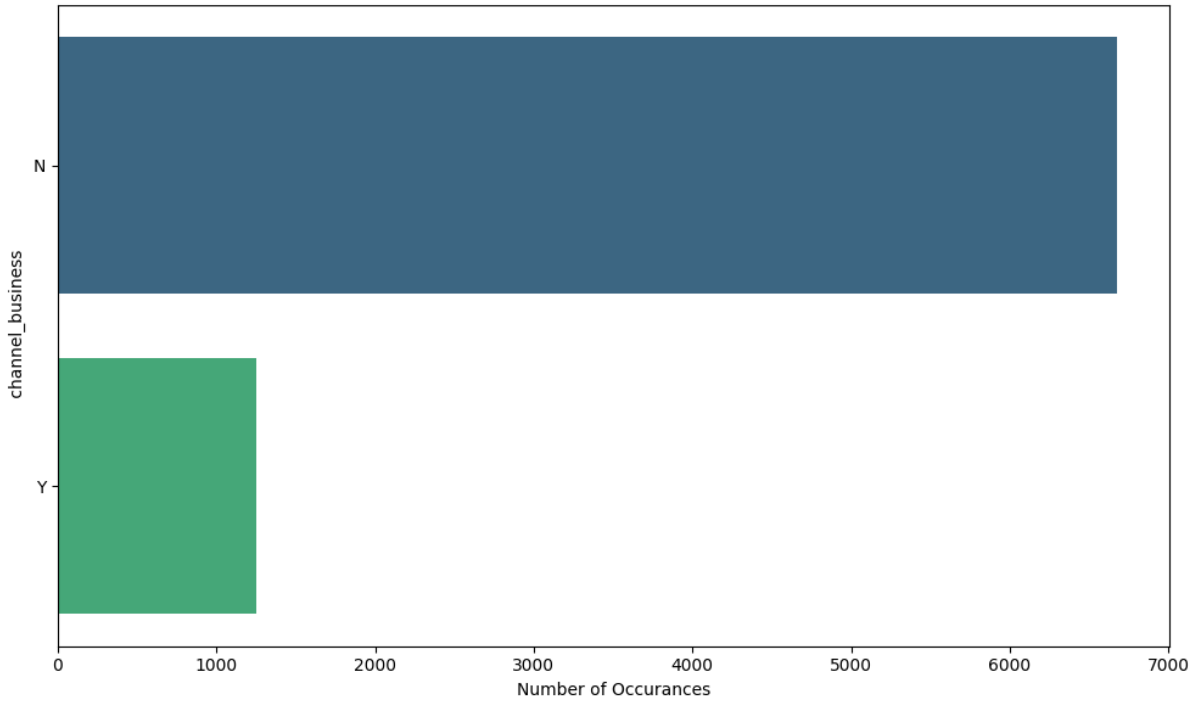

day_tuesday
N
Y
0
1000
2000
3000
4000
5000
6000
Number of Occurances

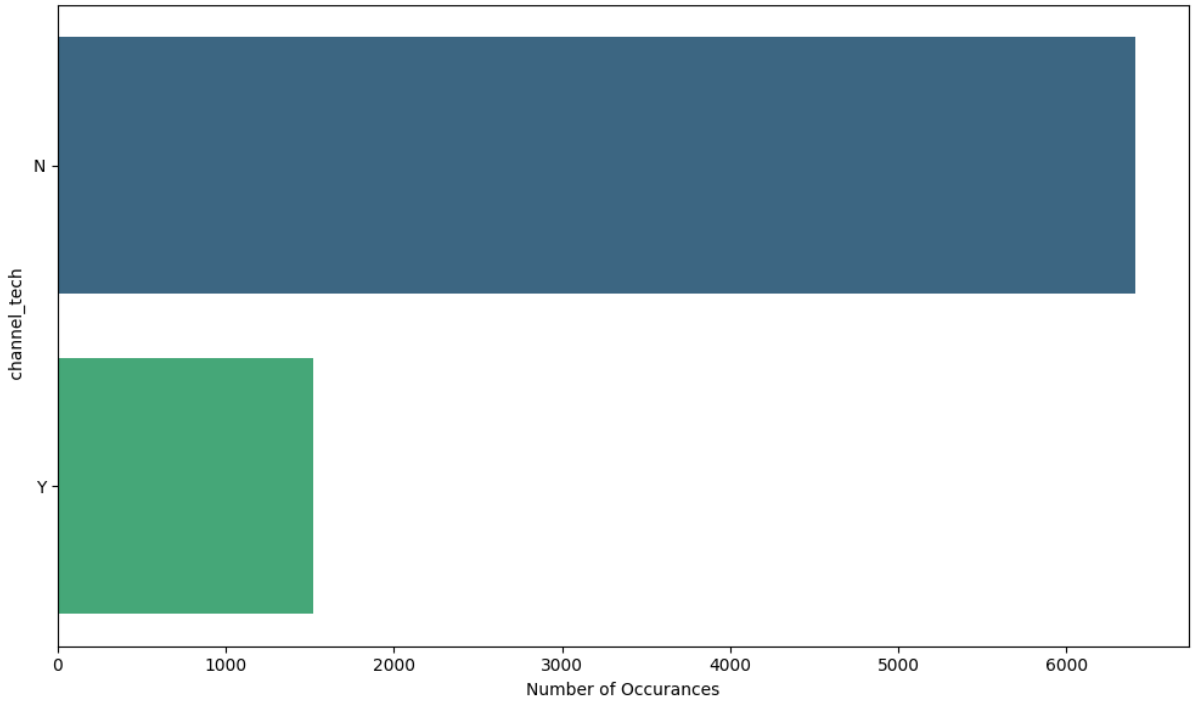

Slightly Popular
Moderately Popular
Popular
Viral
Unpopular
popularity_category
0
500
1000
1500
2000
2500
3000
3500
Number of Occurances

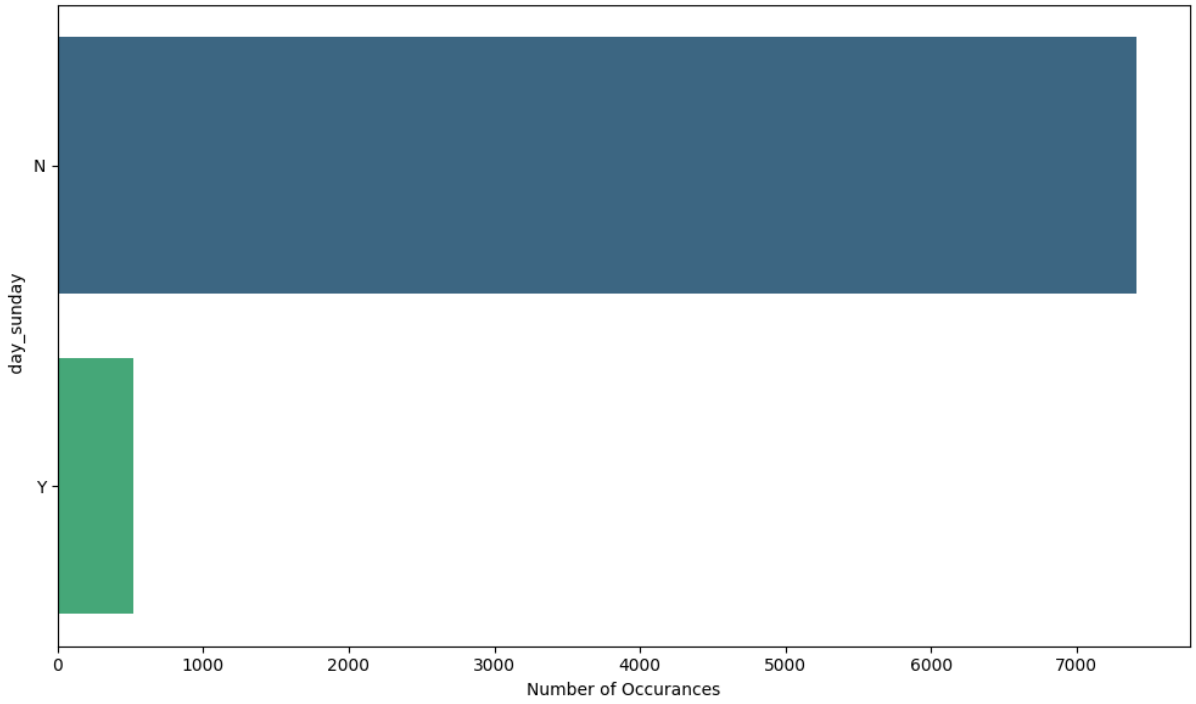

N
Y
channel_lifestyle
0
1000
2000
3000
4000
5000
6000
7000
Number of Occurances

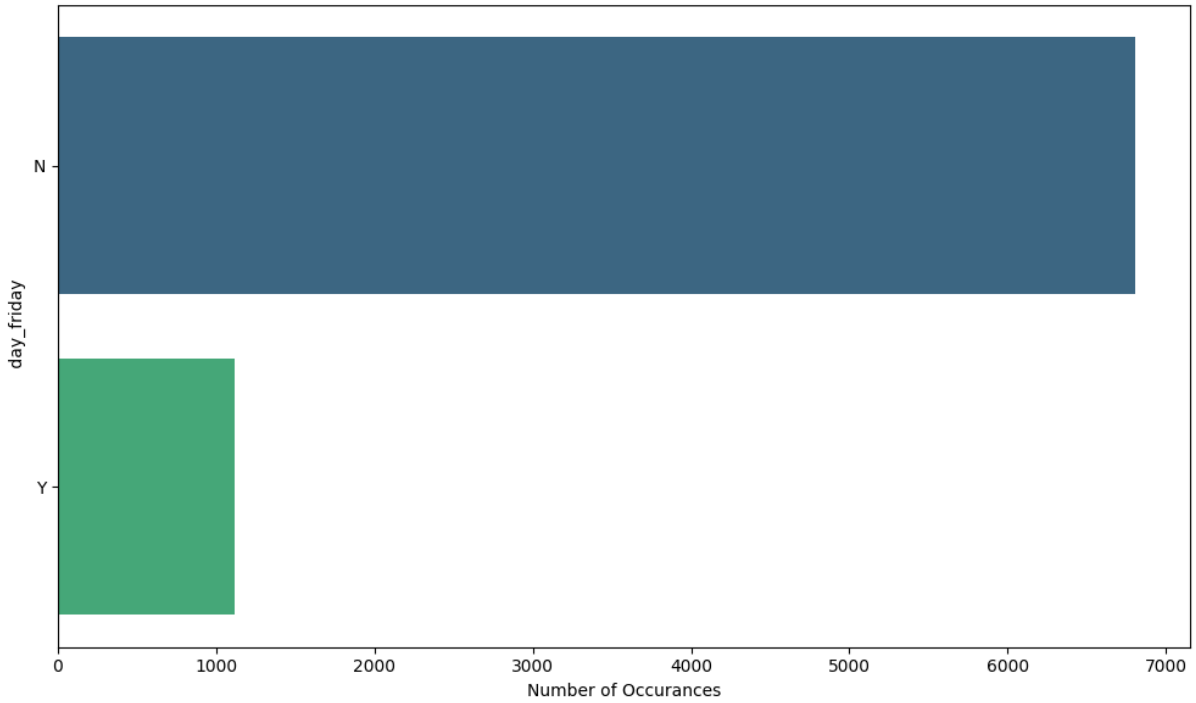

day_thursday
N
Y
0
1000
2000
3000
4000
5000
6000
Number of Occurances


day_monday
N
Y
0
1000
2000
3000
4000
5000
6000
Number of Occurances


channel_business
N
Y
0
1000
2000
3000
4000
5000
6000
7000
Number of Occurances


channel_tech
N
Y
0
1000
2000
3000
4000
5000
6000
Number of Occurances





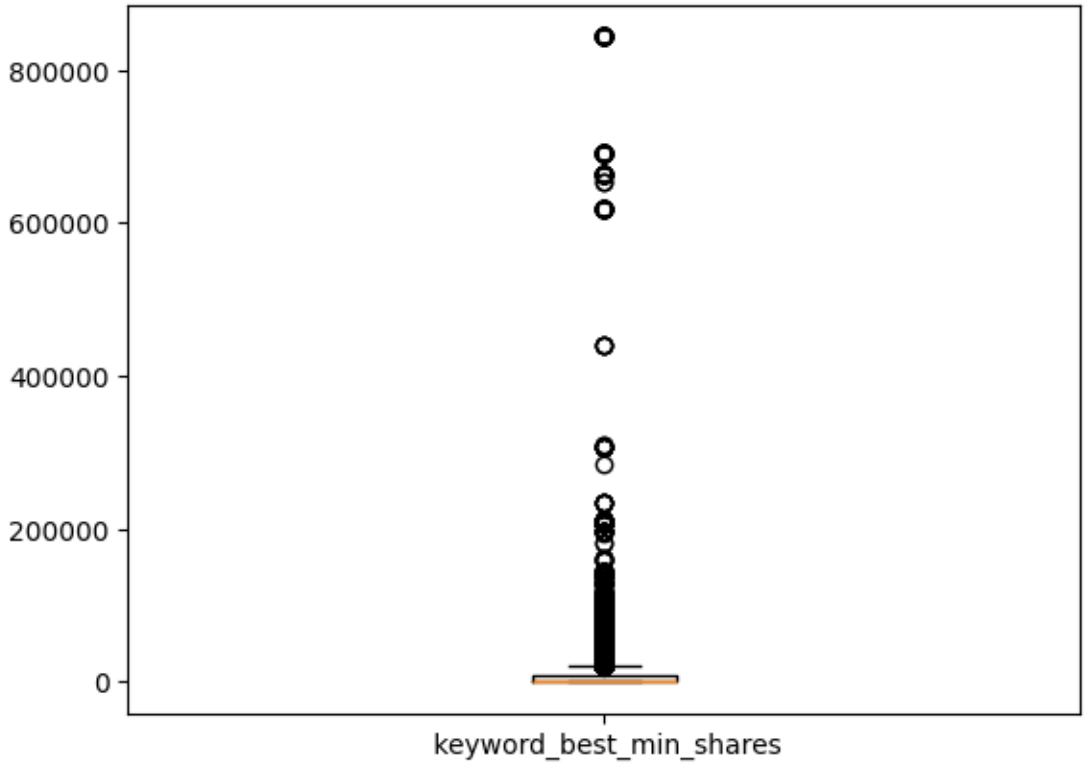
Boxploturi

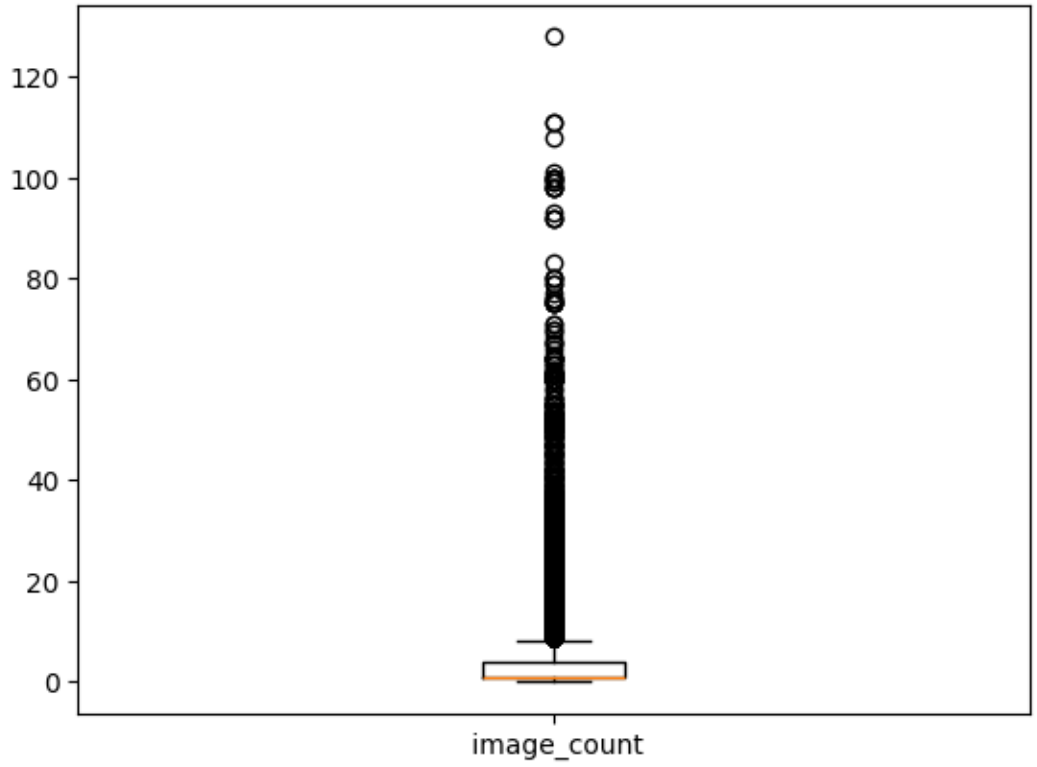
Cutia (IQR) Comprimată: Pentru majoritatea variabilelor continue, box-ul (care reprezintă intervalul interquartilic, între Q1 și Q3) este extrem de turtit și situat în partea de jos a graficului. Acest lucru indică faptul că 50% din articole au caracteristici standard (valori mici și medii), variația fiind minimă pentru majoritatea datelor.

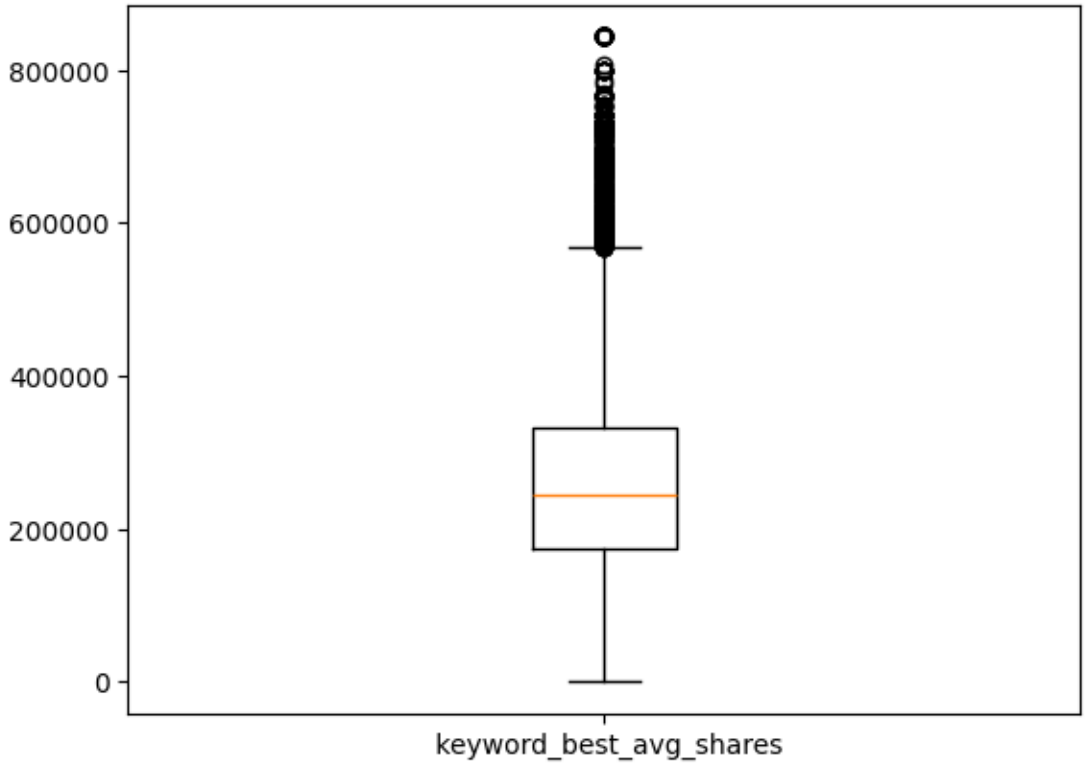
Mustățile (Whiskers) și Valorile Extreme:

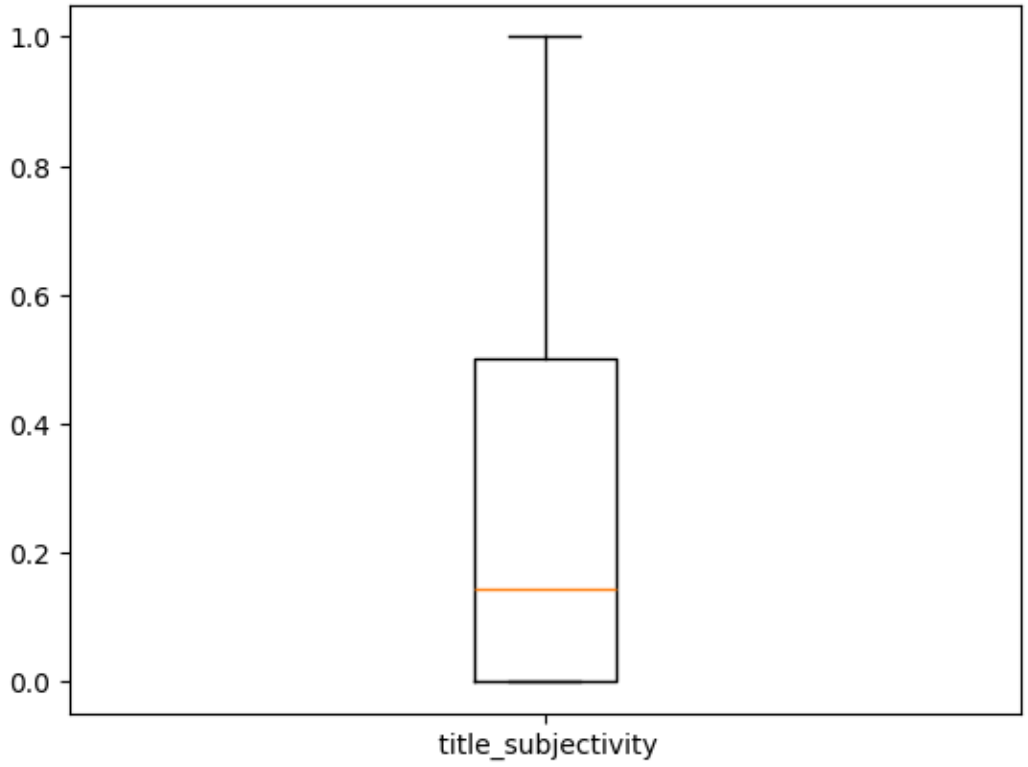
- Se observă un nor dens și continuu de puncte situate mult deasupra "mustății" superioare (Upper Whisker).
- În cazul variabilei content_word_count sau a numărului de link-uri (num_hrefs), punctele de outlier se extind pe o scală uriașă, indicând prezența unor articole atipice care au valori de zeci de ori mai mari decât media (ex: articole foarte lungi sau extrem de virale).
- Aceste puncte confirmă natura de tip "Power Law" (Lege de Putere) a datelor sociale, unde o minoritate de articole acumulează valori extreme.

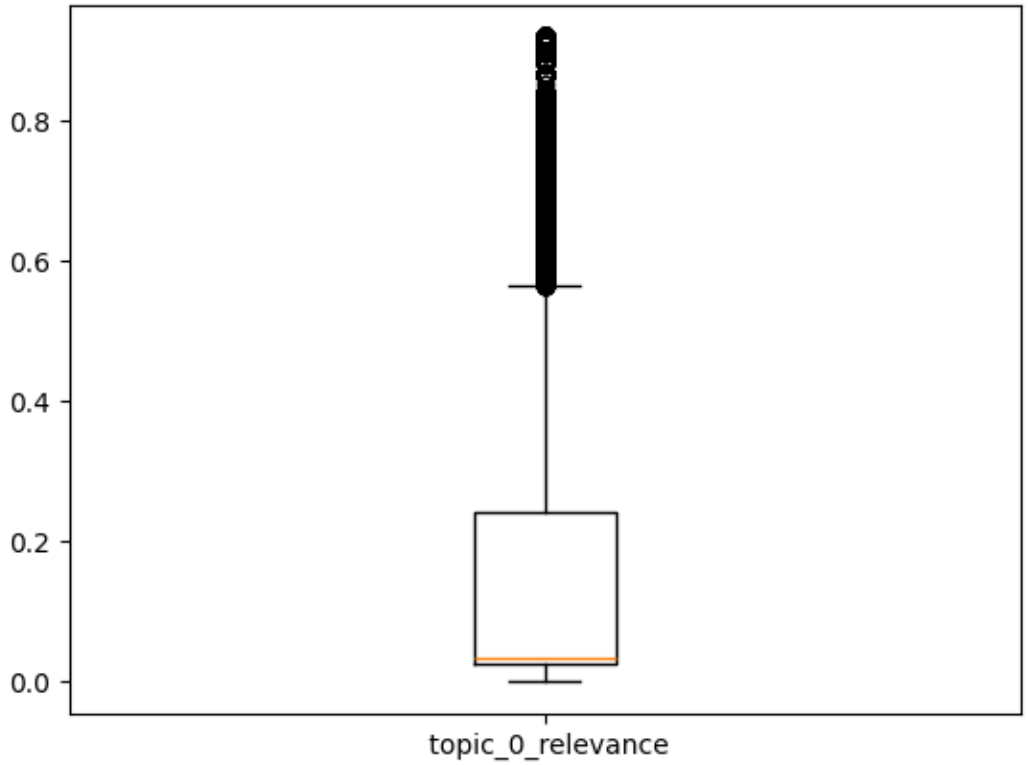
Concluzie vizuală: Distribuția datelor este extrem de asimetrică (highly skewed-right). Prezența masivă a punctelor negre (outlieri) deasupra limitei superioare justifică necesitatea utilizării unor metode robuste de scalare și aplicarea unor praguri de filtrare a outlierilor mai relaxate (percentilele 10-90) pentru a nu elimina informația despre articolele de top.

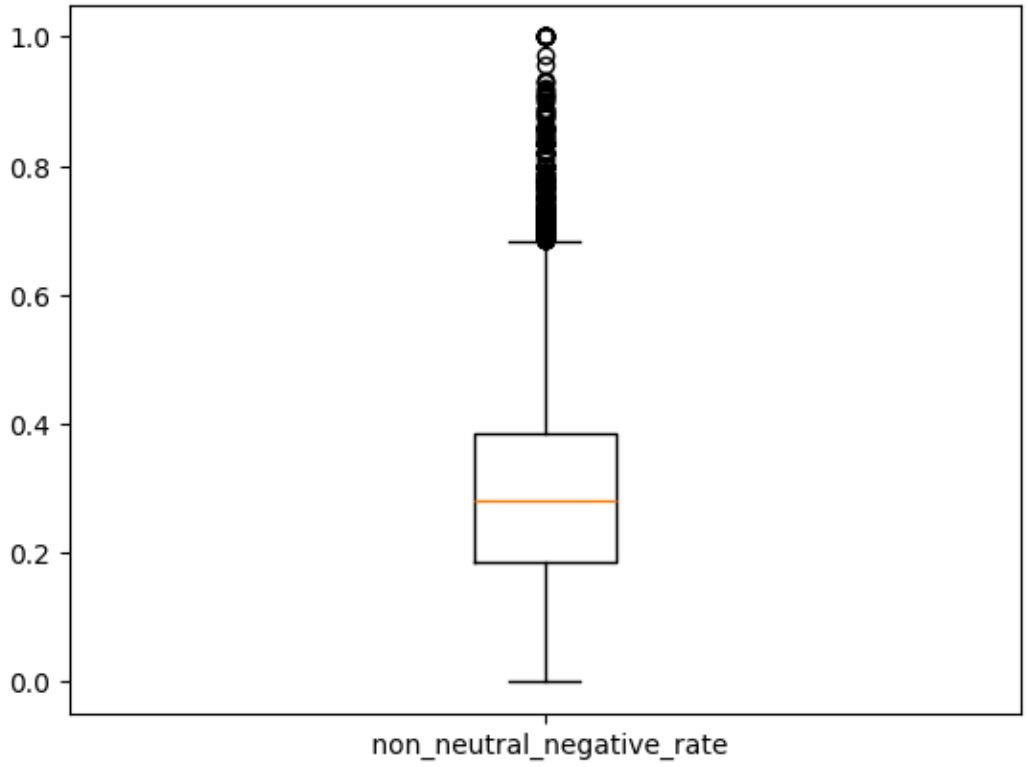

800000
600000
400000
200000
0
keyword_best_min_shares

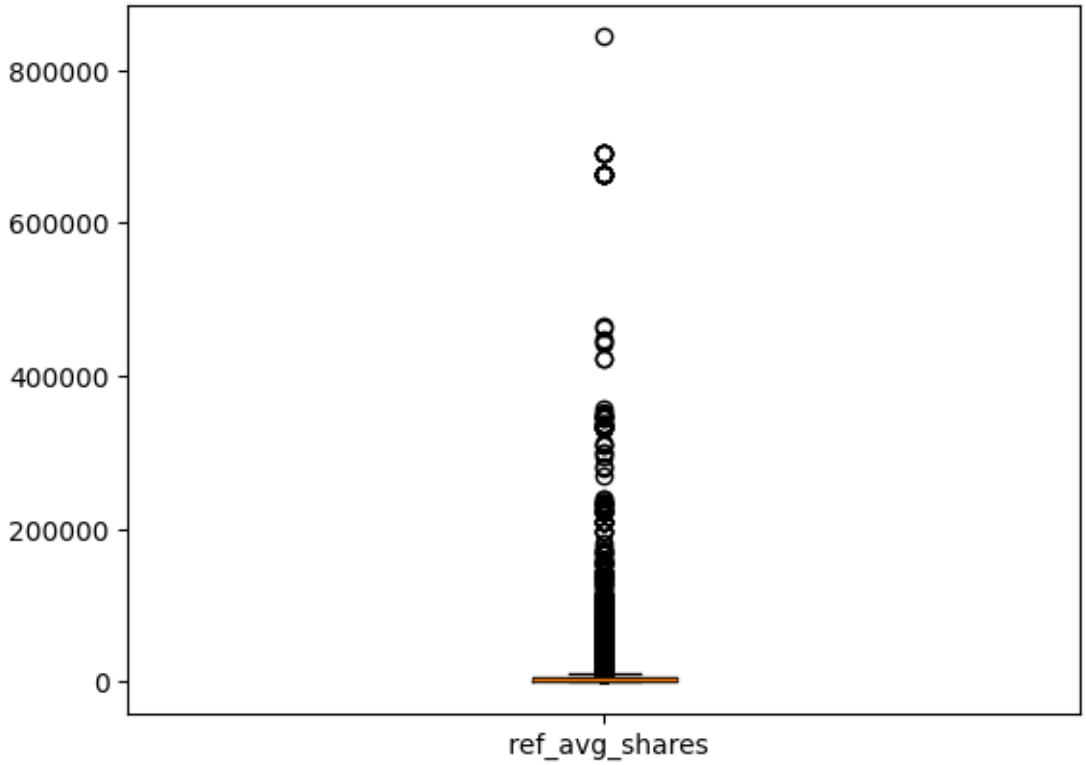

120
100
80
60
40
20
0
image_count

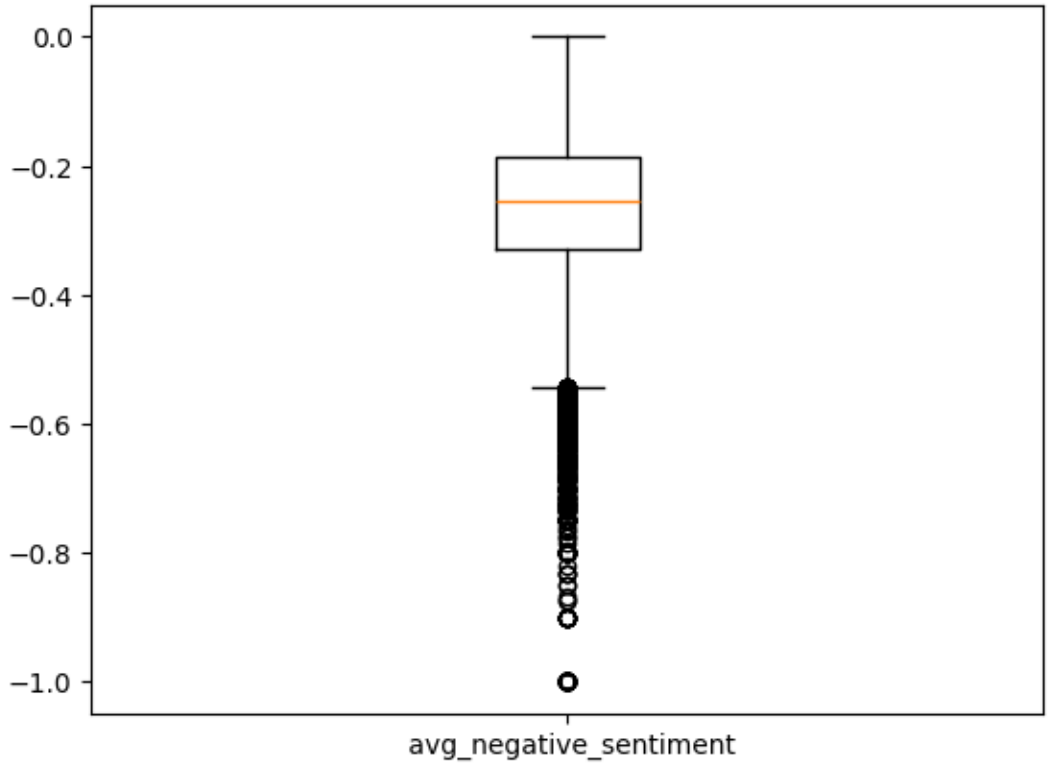

800000
600000
400000
200000
0
keyword_best_avg_shares

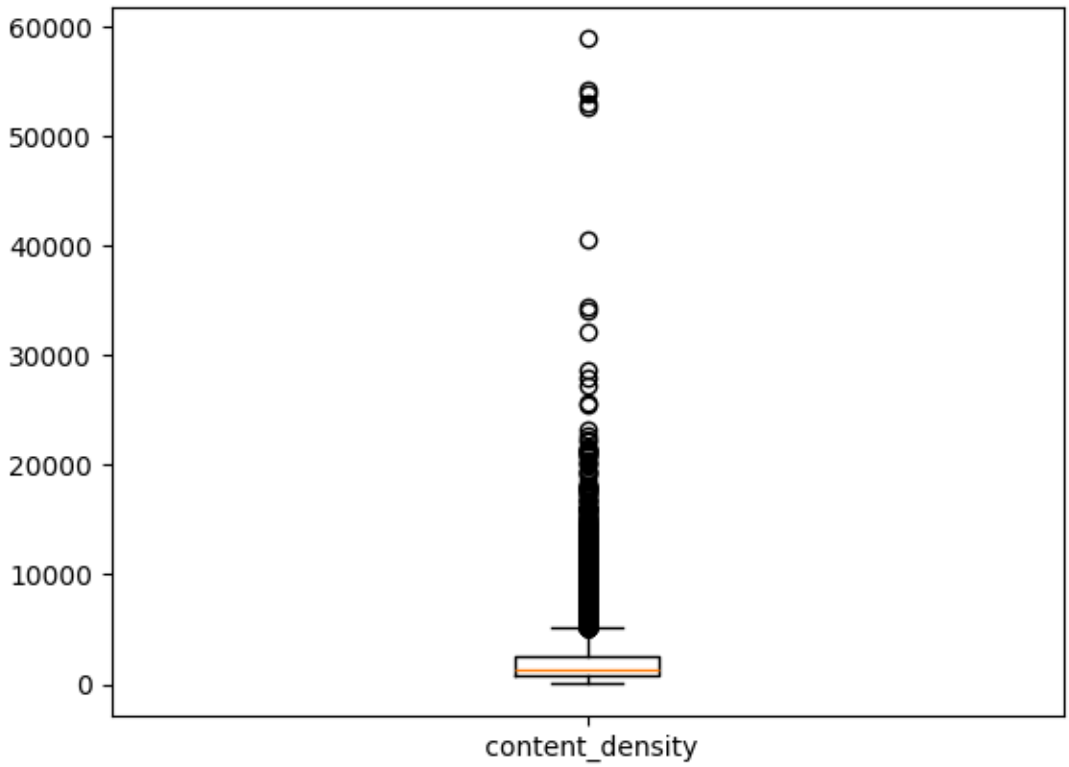

1.0
0.8
0.6
0.4
0.2
0.0
title_subjectivity

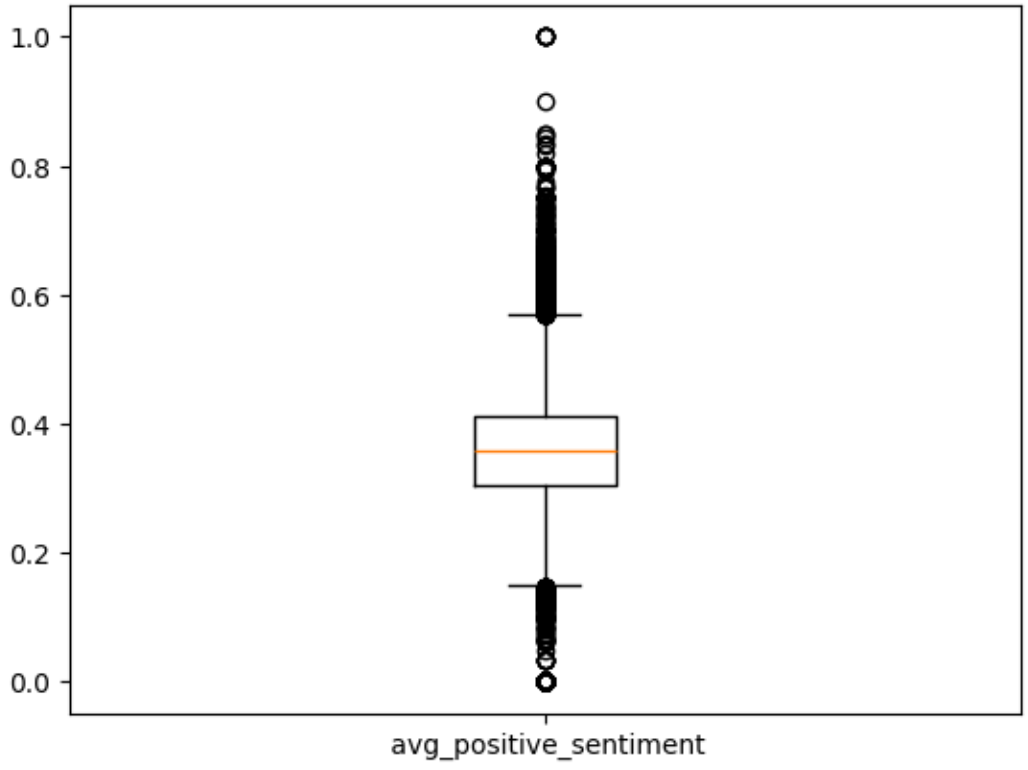

0.8
0.6
0.4
0.2
0.0
topic_0_relevance

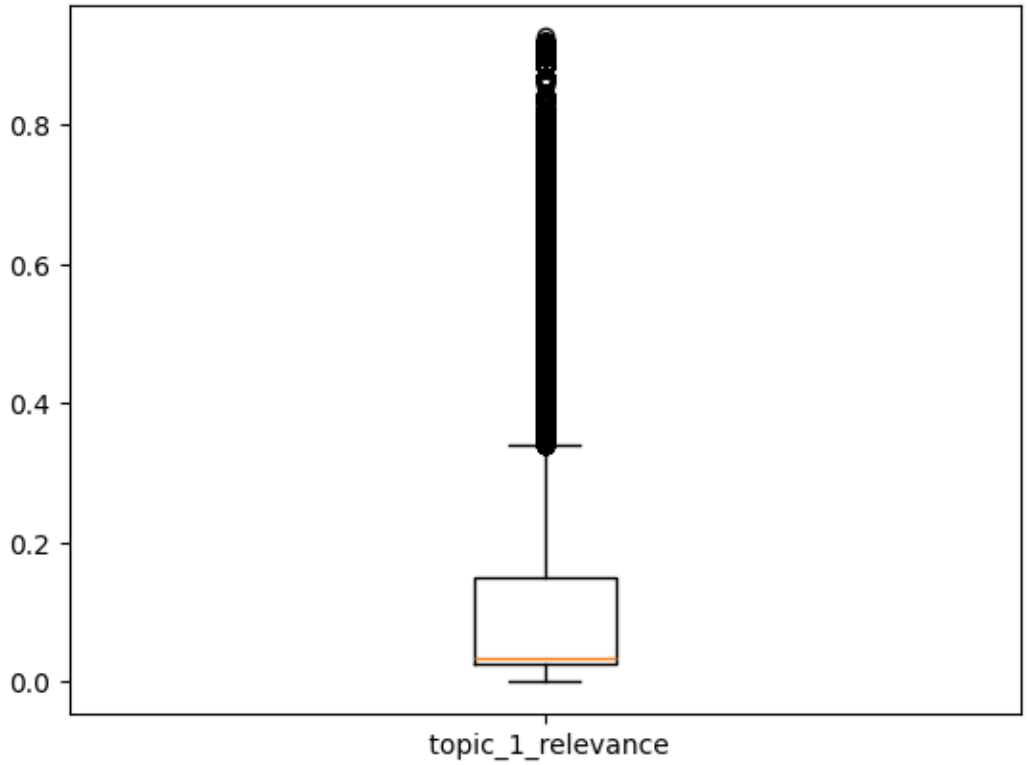

1.0
0.8
0.6
0.4
0.2
0.0
non_neutral_negative_rate

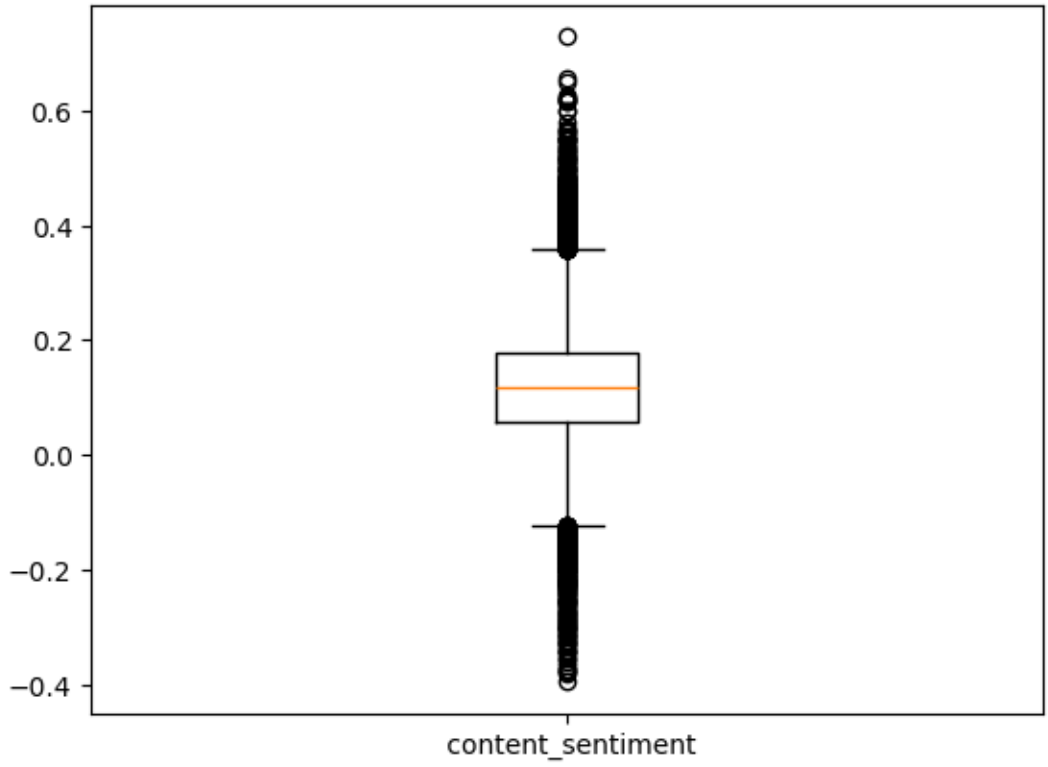

800000
600000
400000
200000
0
ref_avg_shares

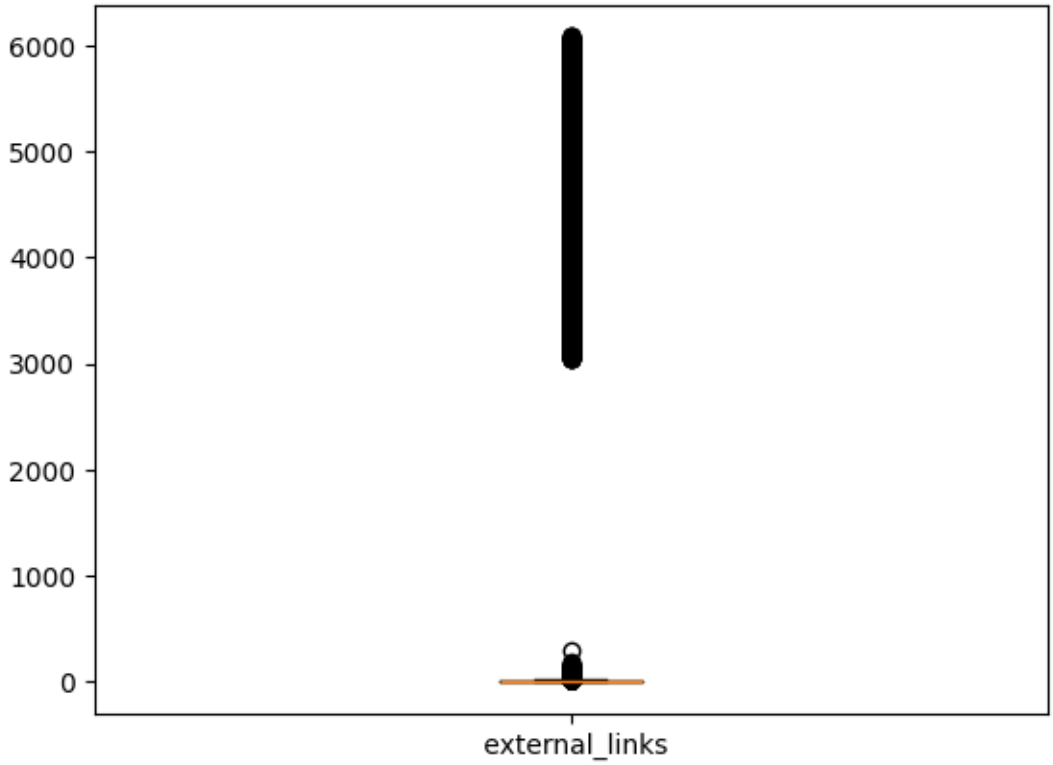

0.0
−0.2
−0.4
−0.6
−0.8
−1.0
avg_negative_sentiment

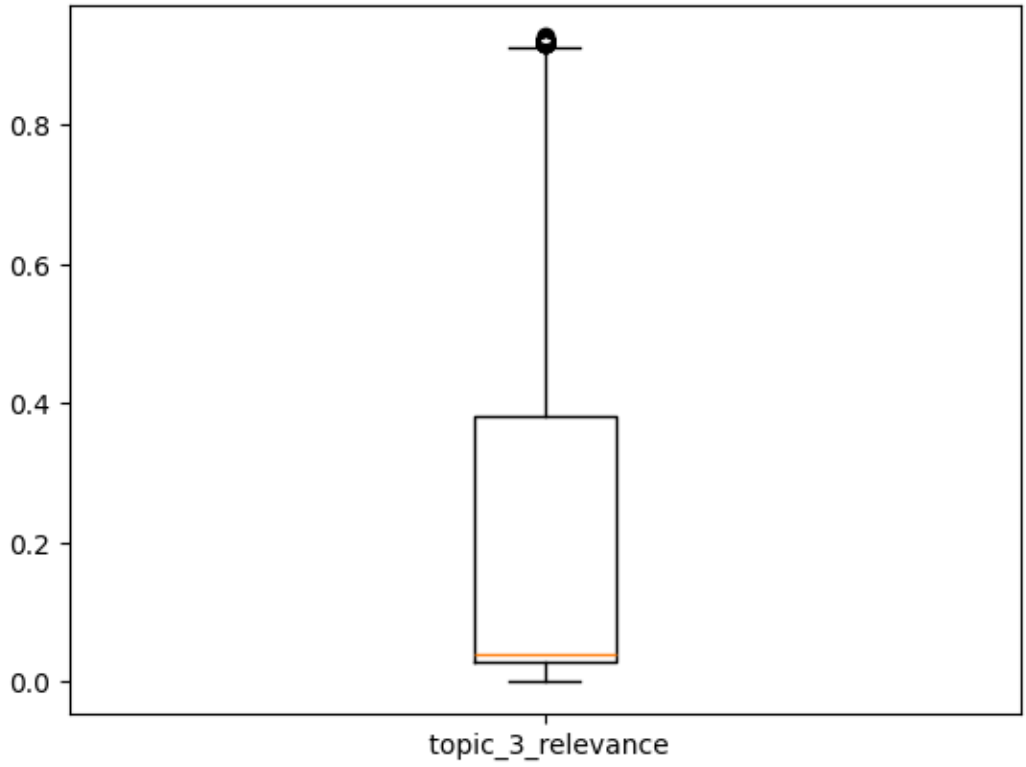

60000
50000
40000
30000
20000
10000
0
content_density

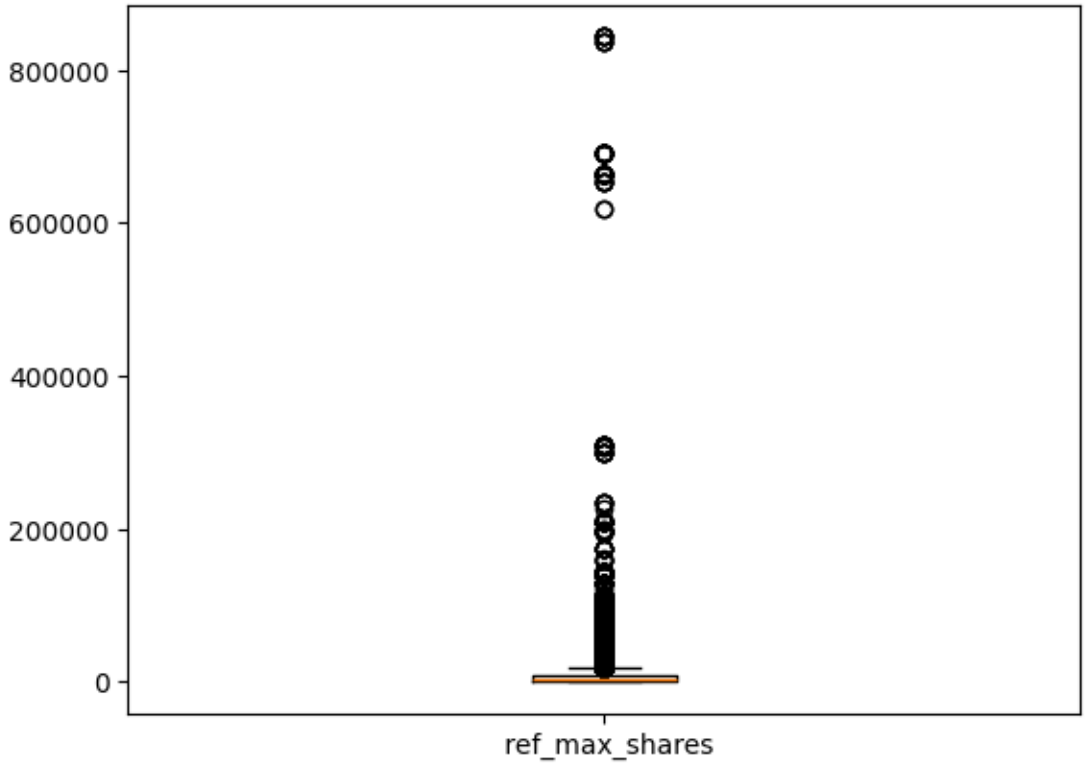

1.0
0.8
0.6
0.4
0.2
0.0
avg_positive_sentiment

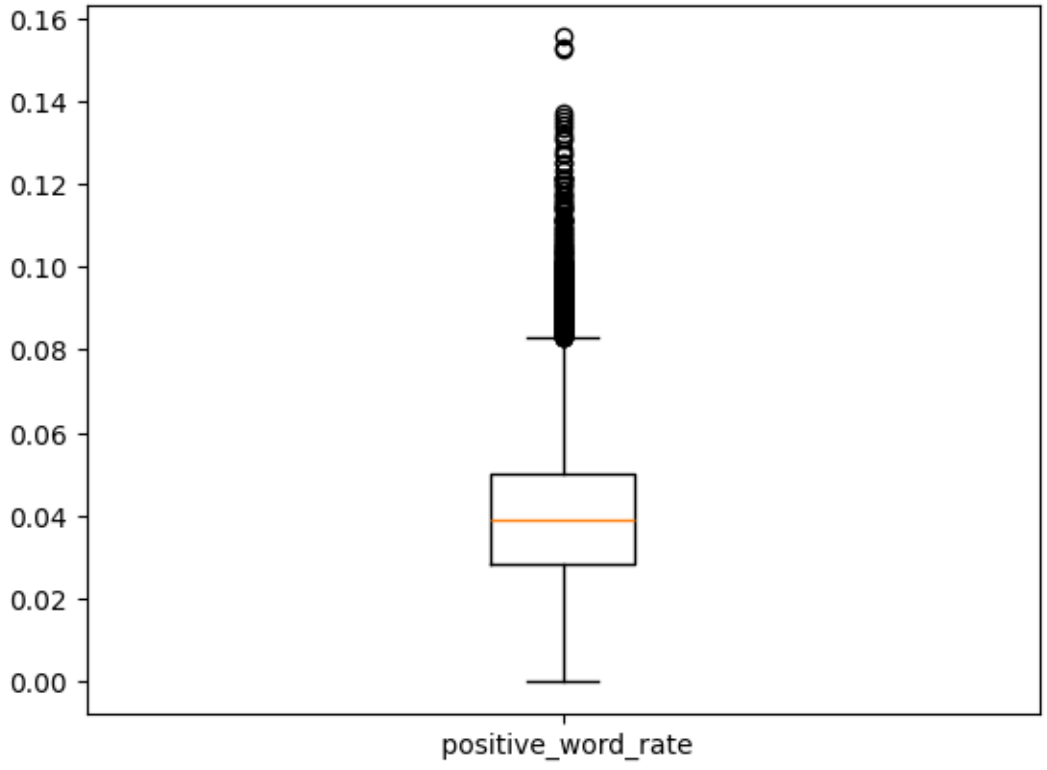

0.8
0.6
0.4
0.2
0.0
topic_1_relevance

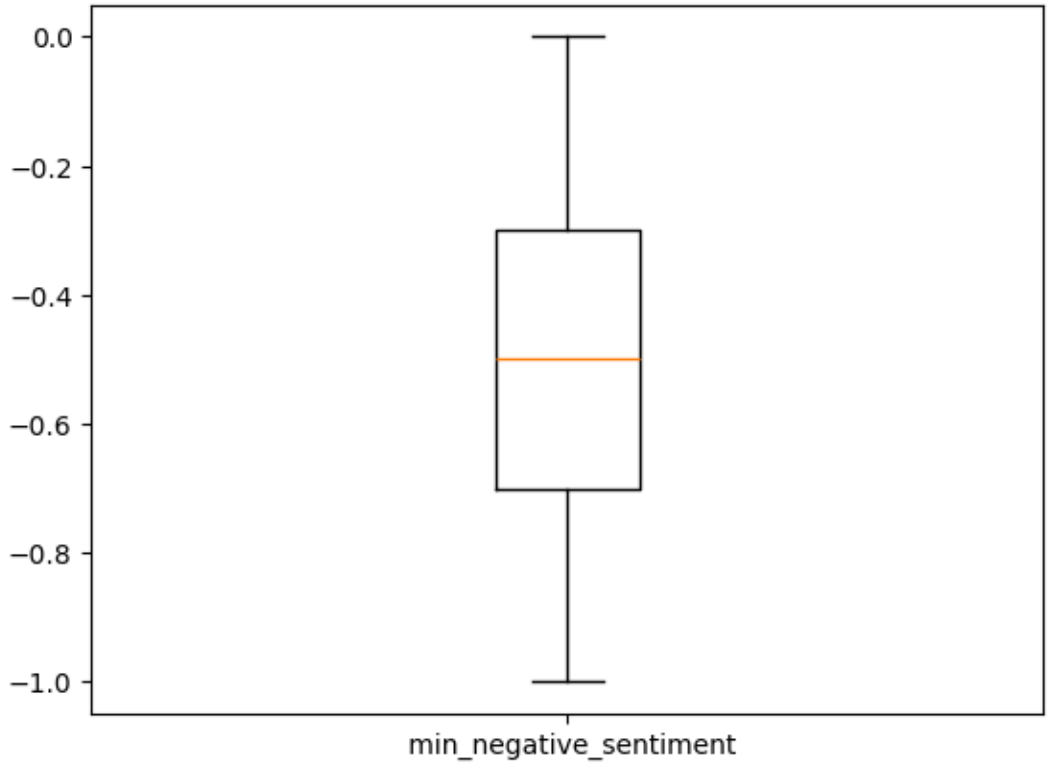

0.6
0.4
0.2
0.0
−0.2
−0.4
content_sentiment

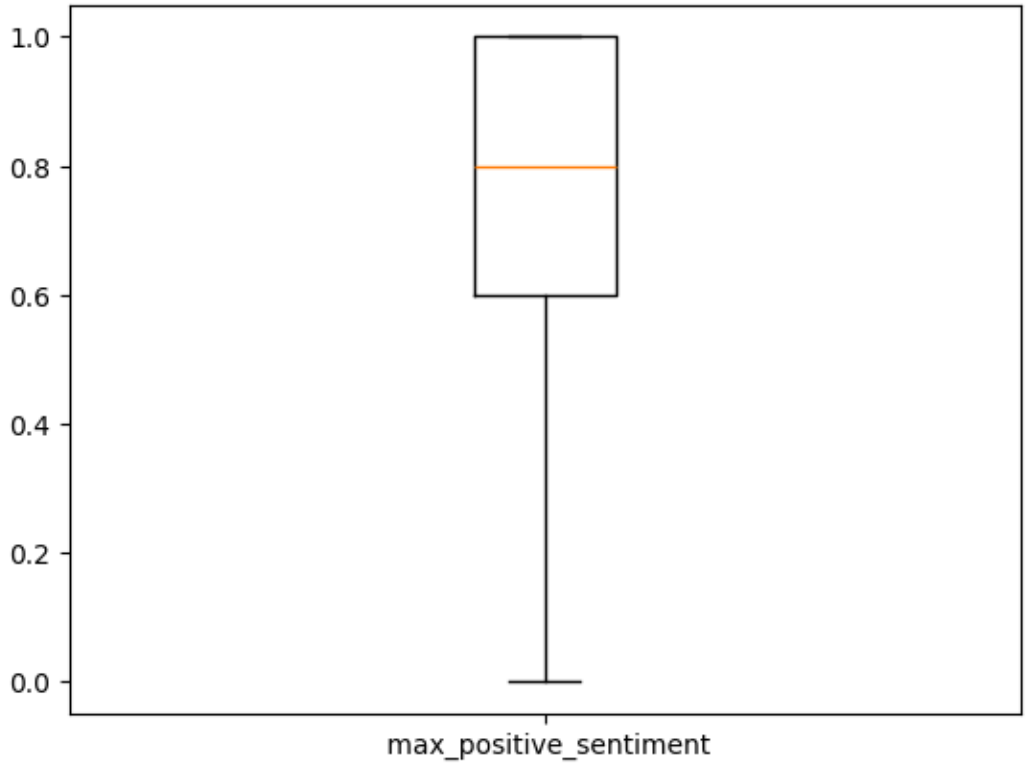

6000
5000
4000
3000
2000
1000
0
external_links

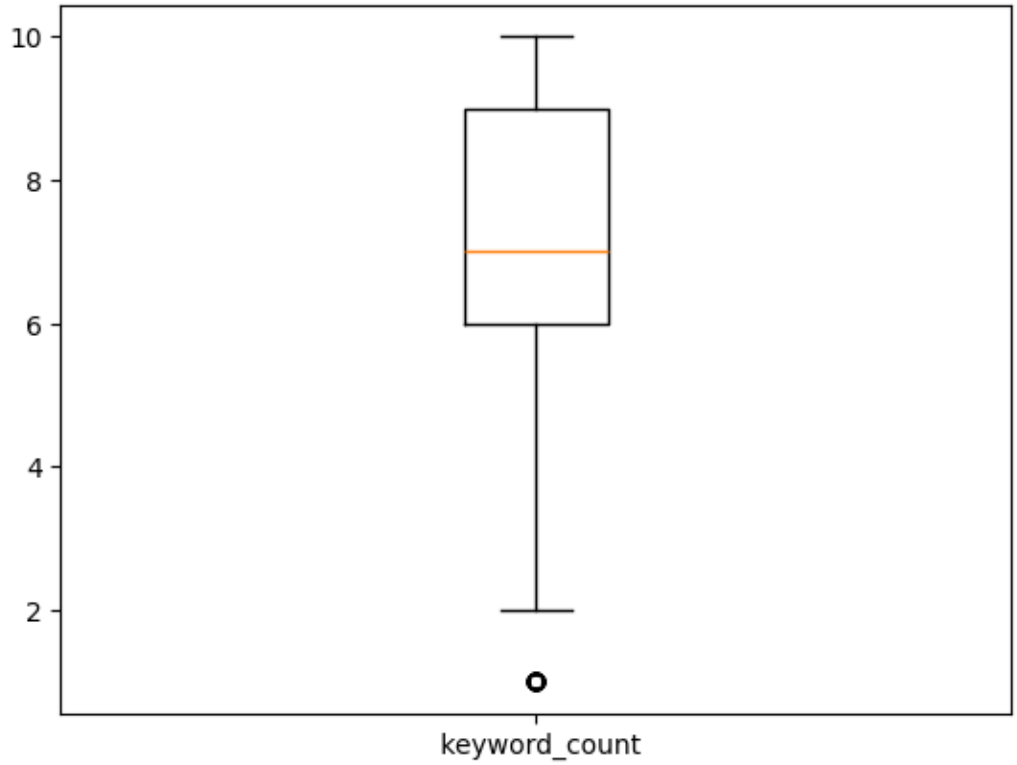

0.8
0.6
0.4
0.2
0.0
topic_3_relevance

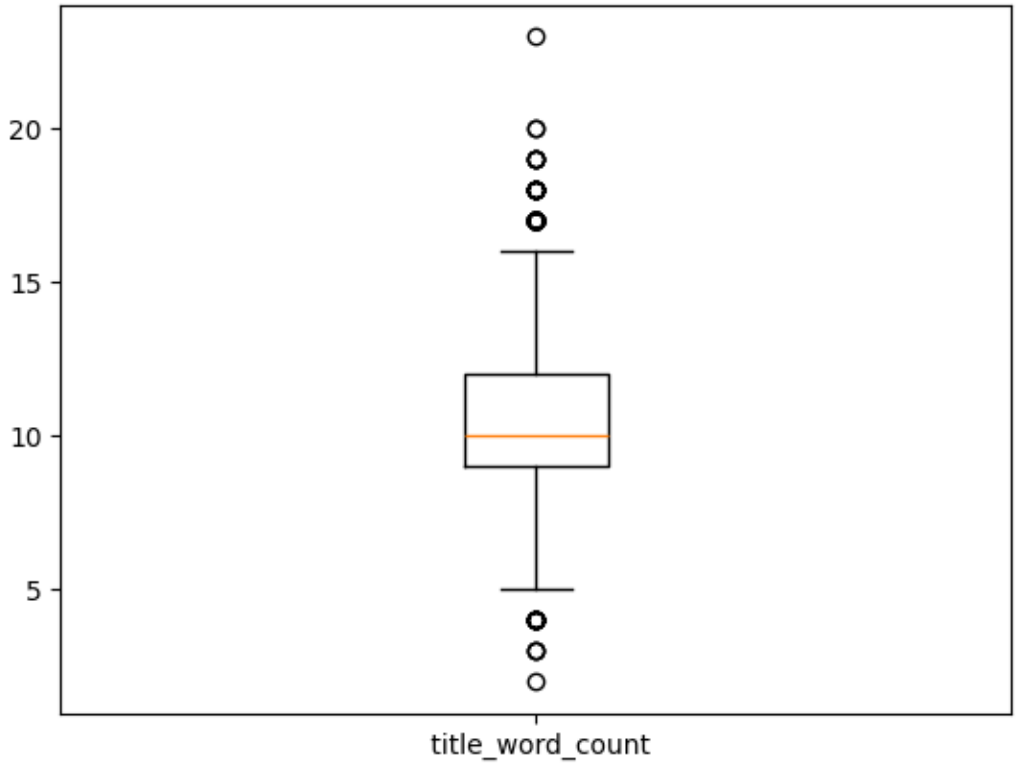

800000
600000
400000
200000
0
ref_max_shares

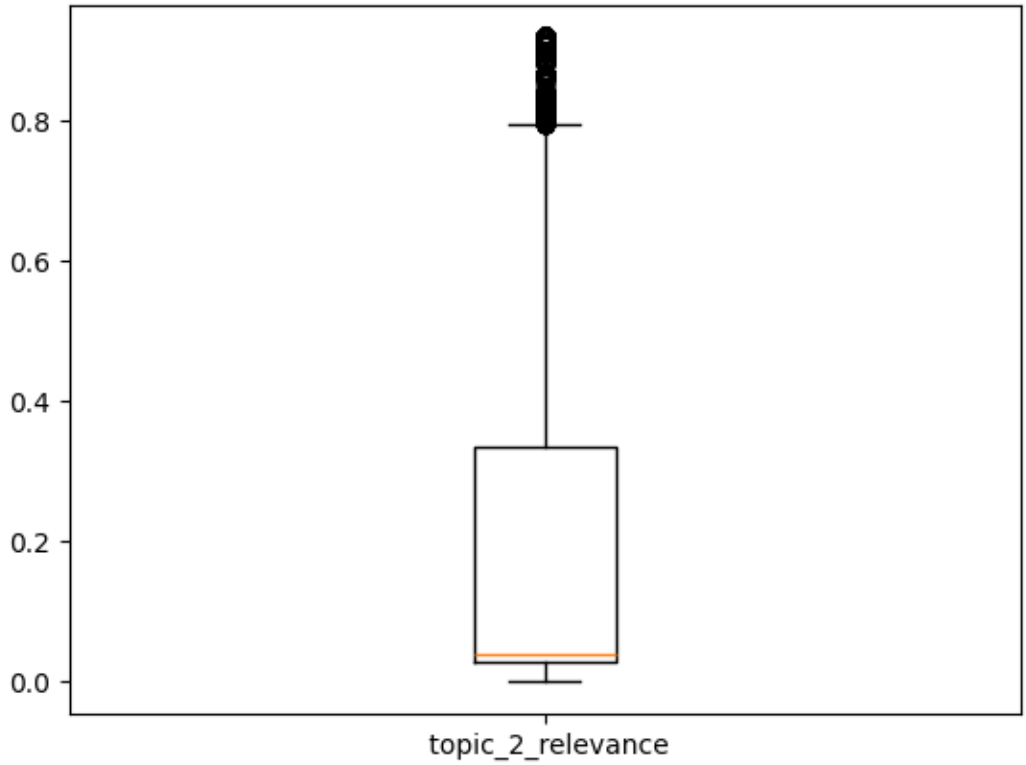

0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
positive_word_rate

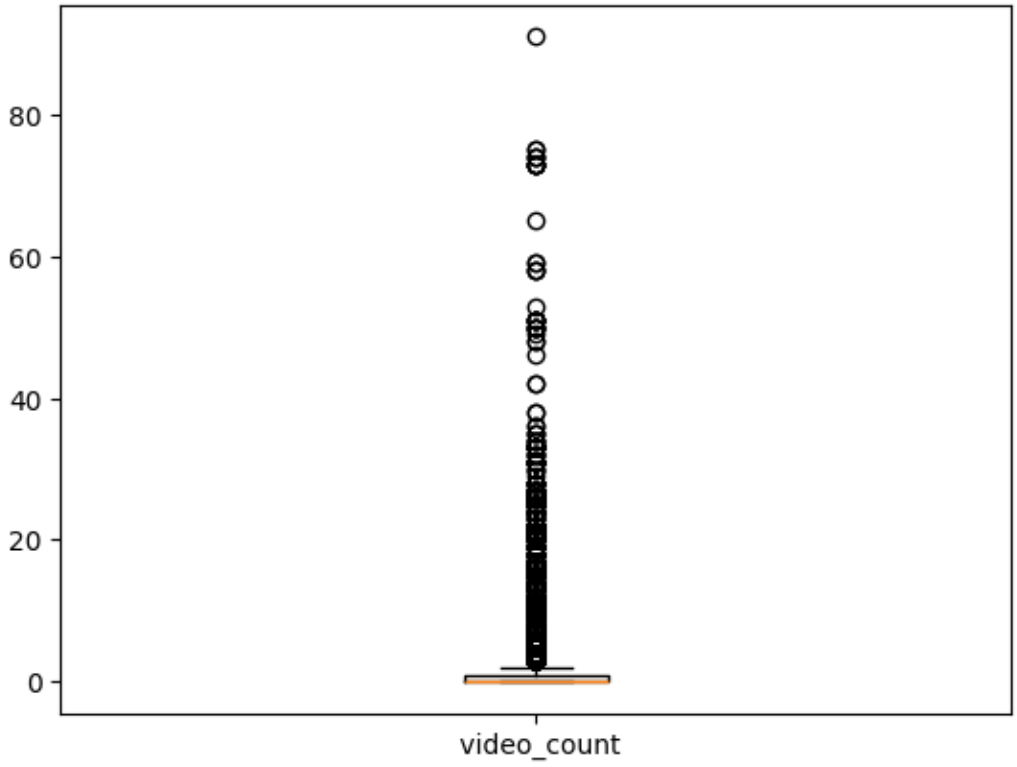

0.0
−0.2
−0.4
−0.6
−0.8
−1.0
min_negative_sentiment

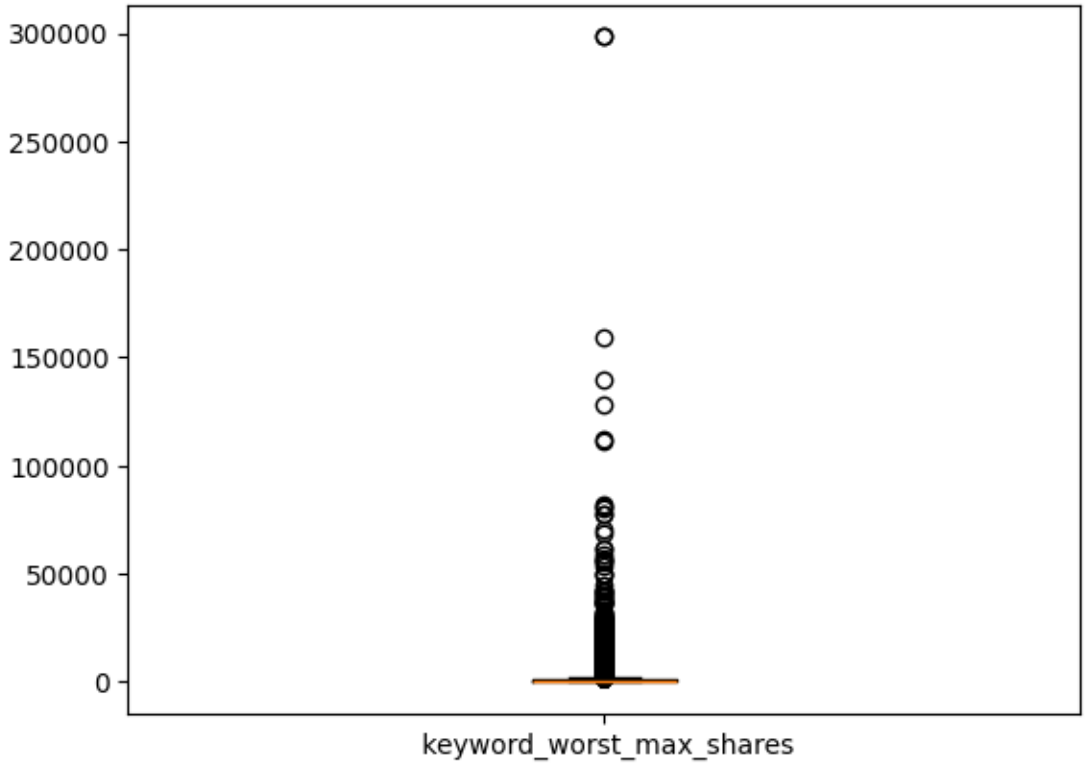

1.0
0.8
0.6
0.4
0.2
0.0
max_positive_sentiment

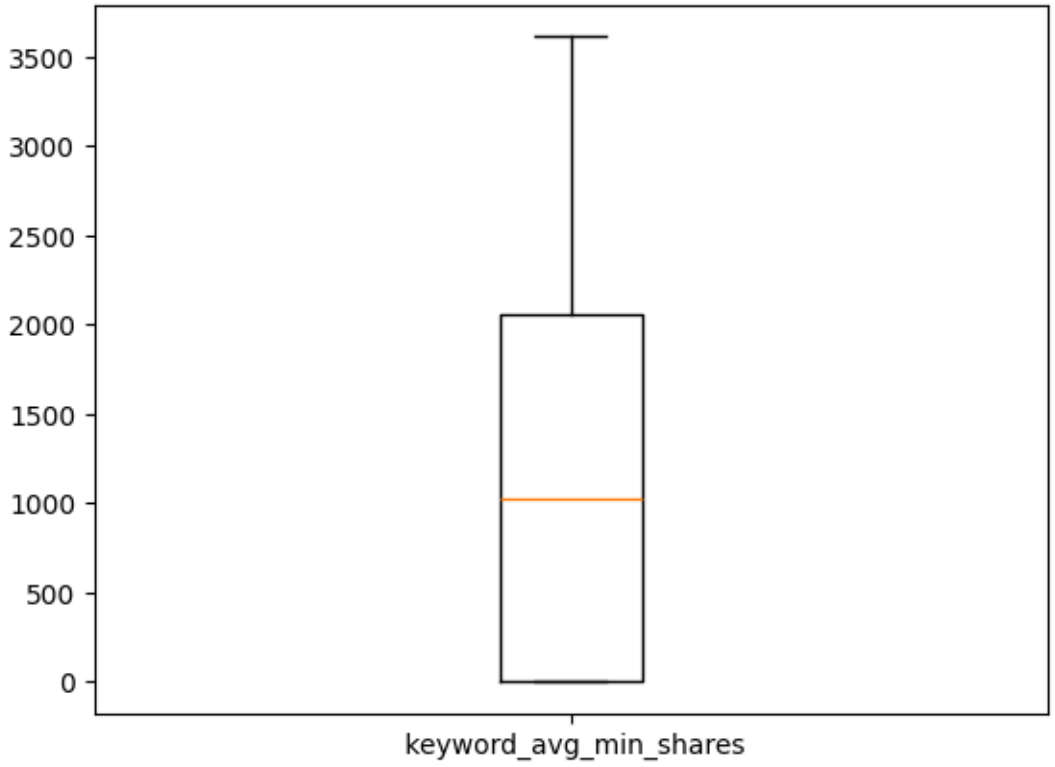

10
8
6
4
2
keyword_count

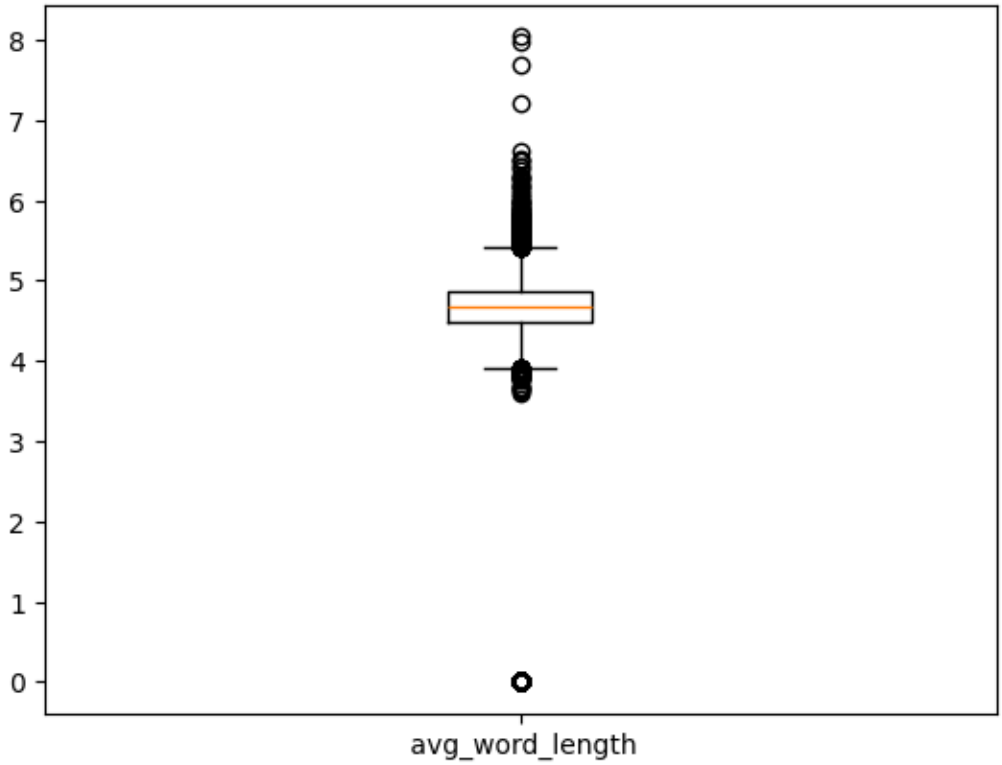

20
15
10
5
title_word_count

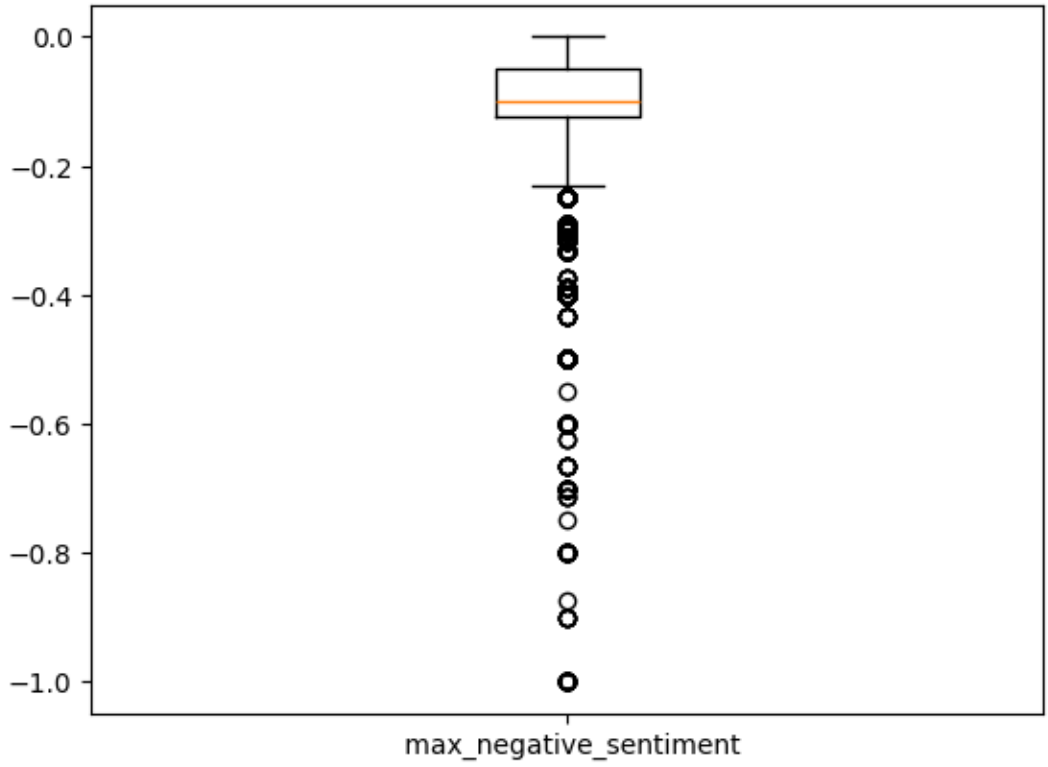

0.8
0.6
0.4
0.2
0.0
topic_2_relevance

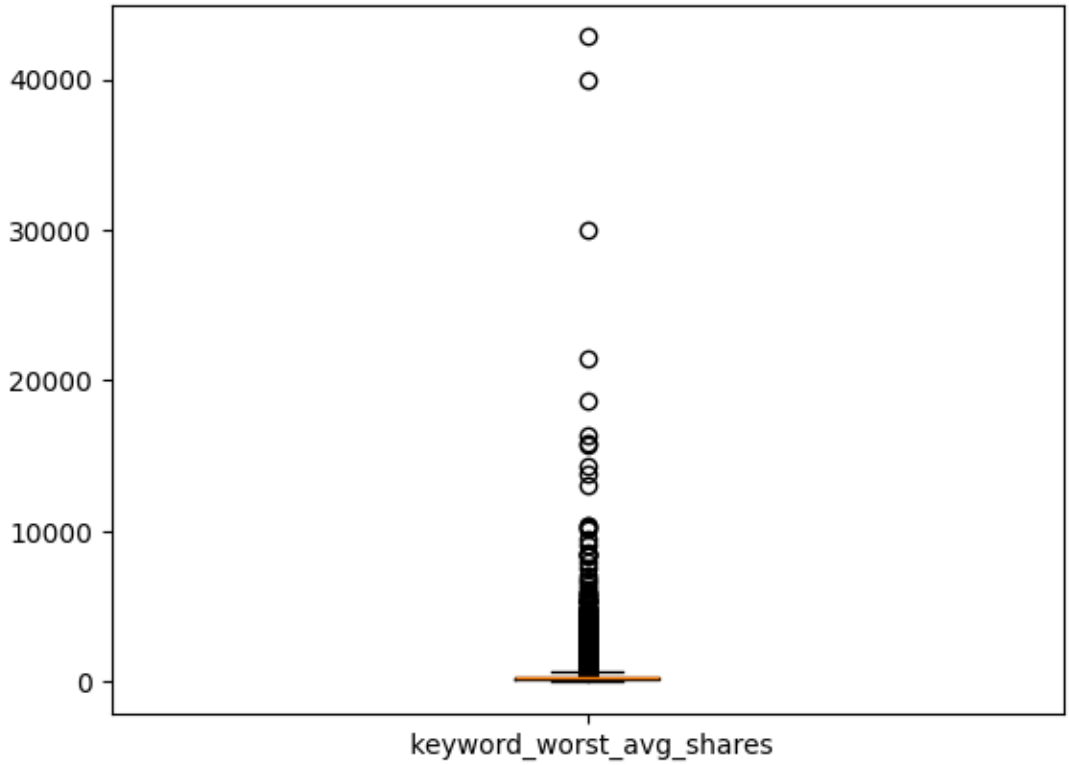

80
60
40
20
0
video_count

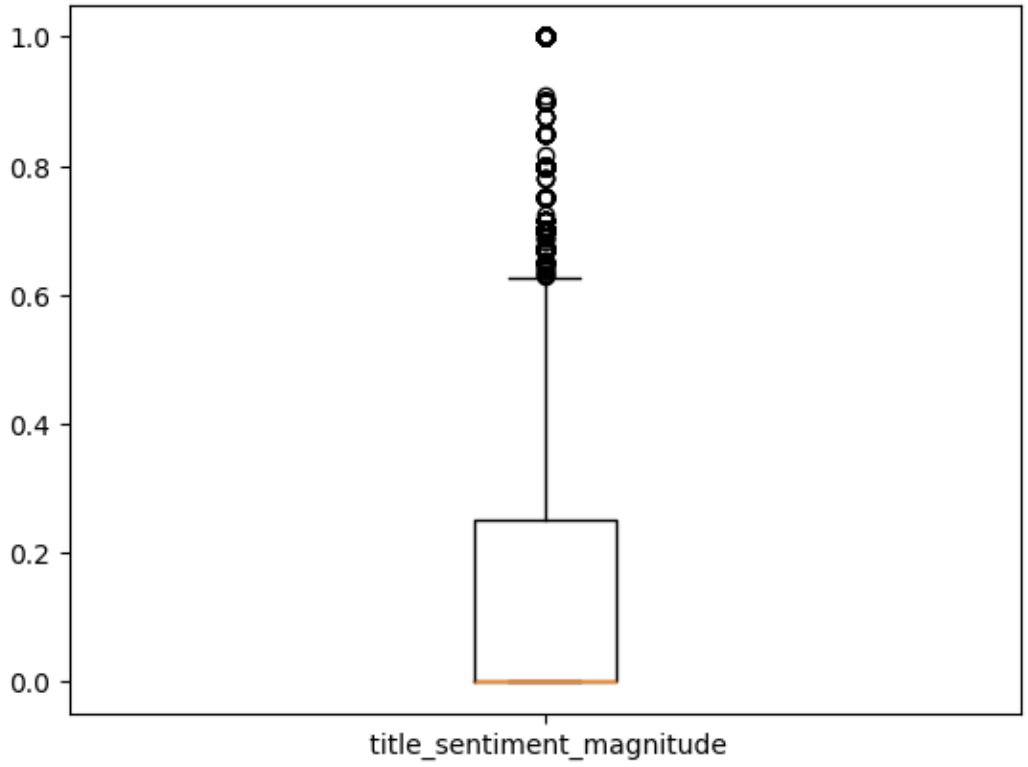

300000
250000
200000
150000
100000
50000
0
keyword_worst_max_shares

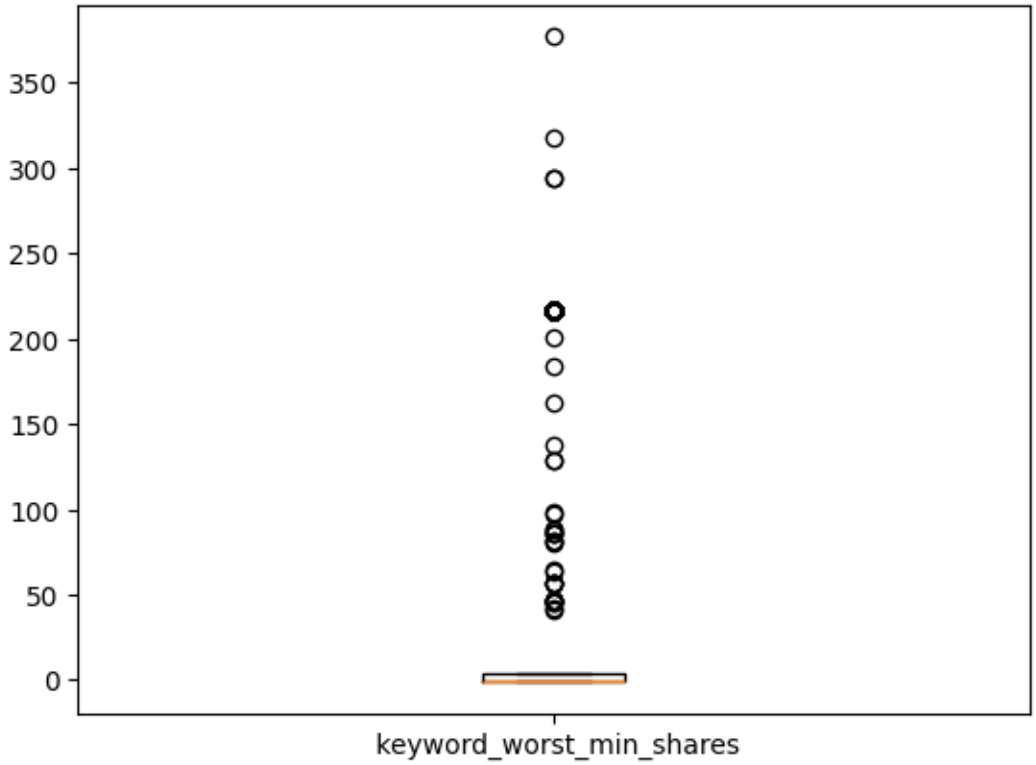

3500
3000
2500
2000
1500
1000
500
0
keyword_avg_min_shares

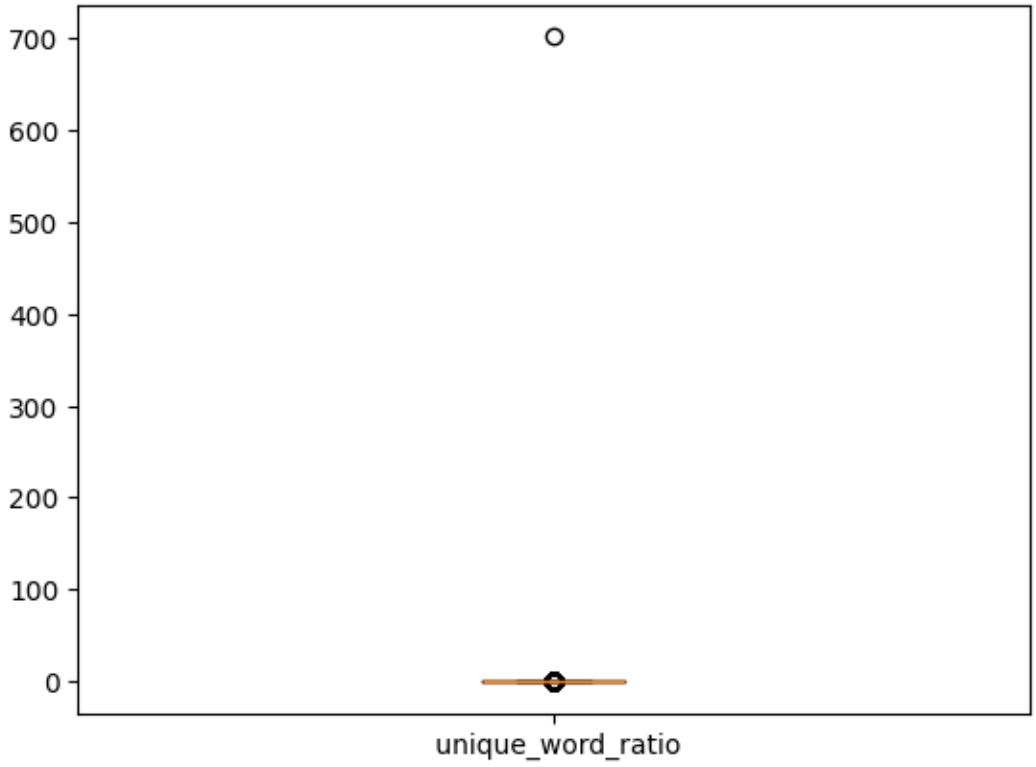

8
7
6
5
4
3
2
1
0
avg_word_length

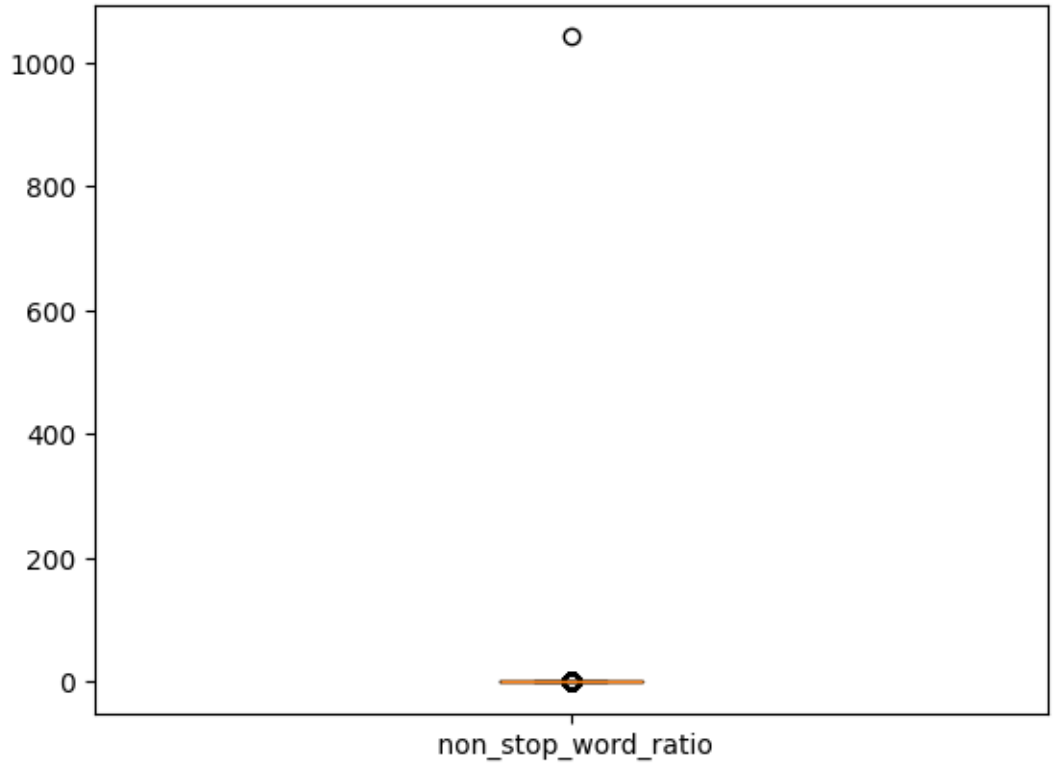

0.0
−0.2
−0.4
−0.6
−0.8
−1.0
max_negative_sentiment

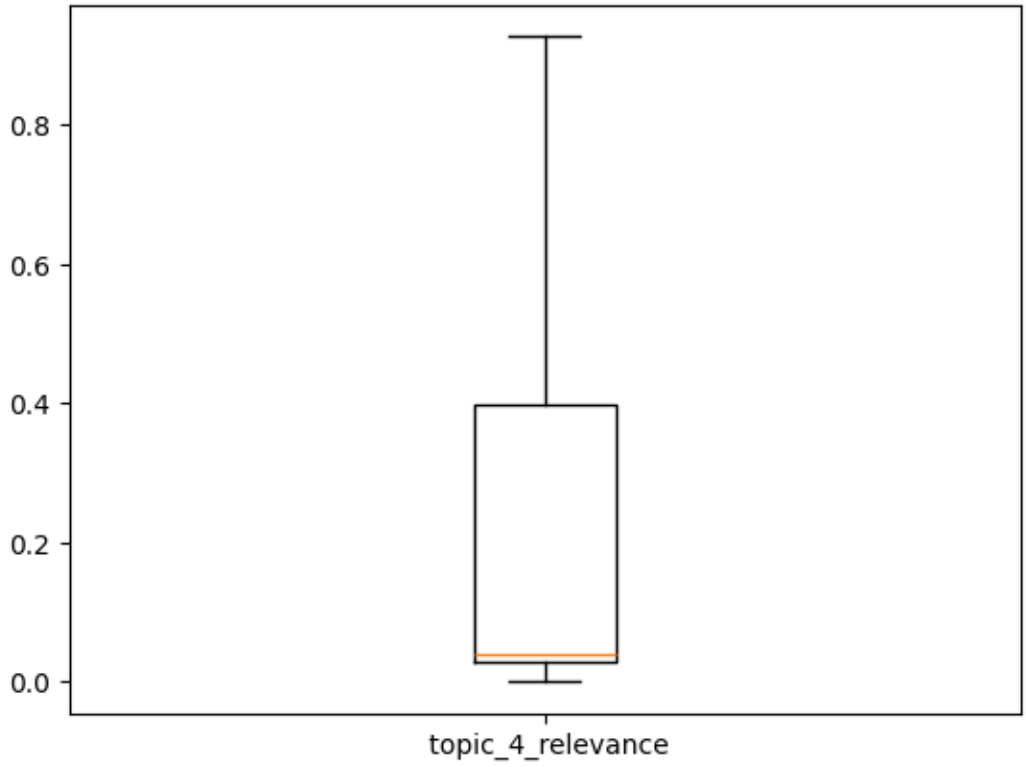

40000
30000
20000
10000
0
keyword_worst_avg_shares

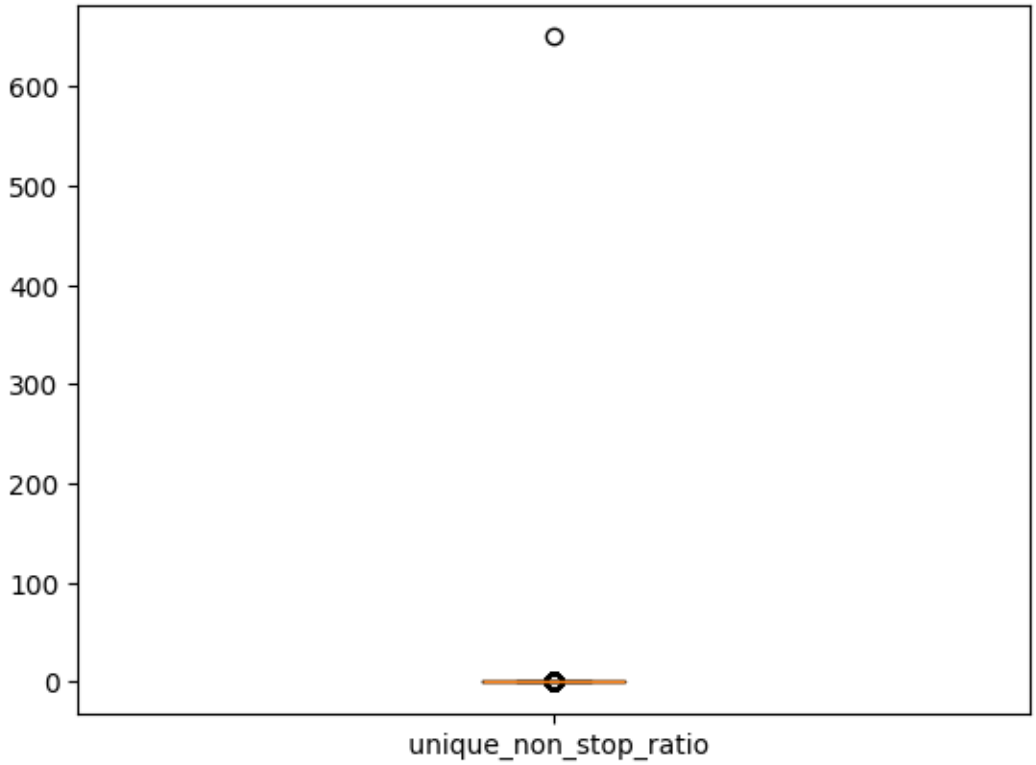

1.0
0.8
0.6
0.4
0.2
0.0
title_sentiment_magnitude

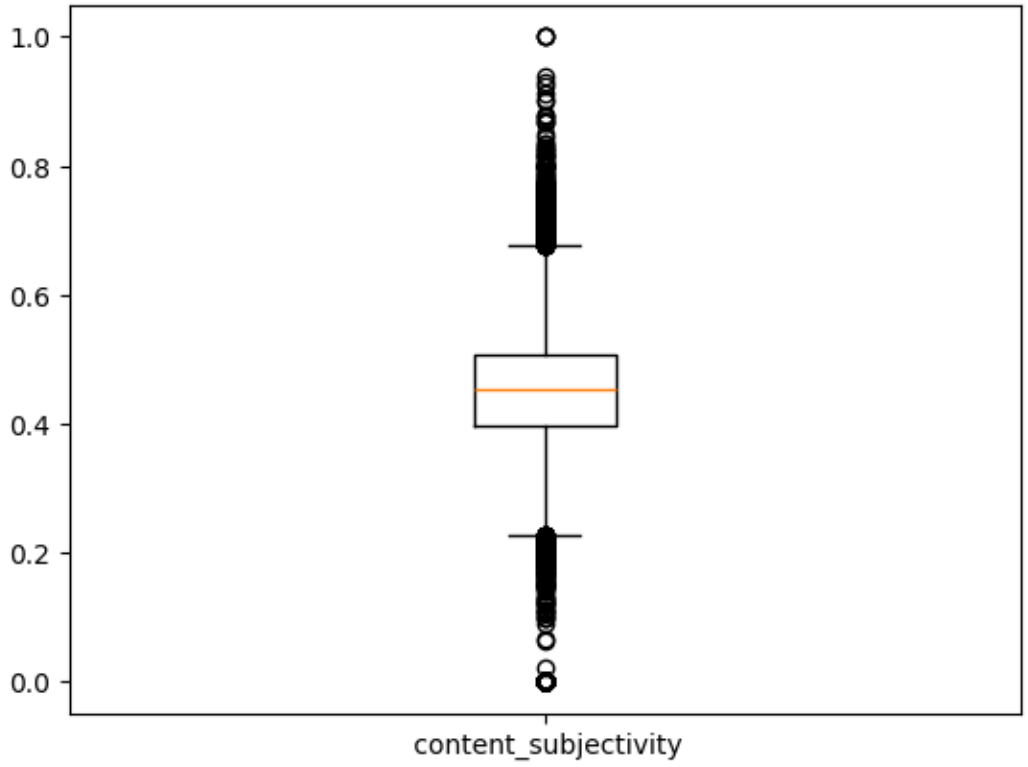

350
300
250
200
150
100
50
0
keyword_worst_min_shares

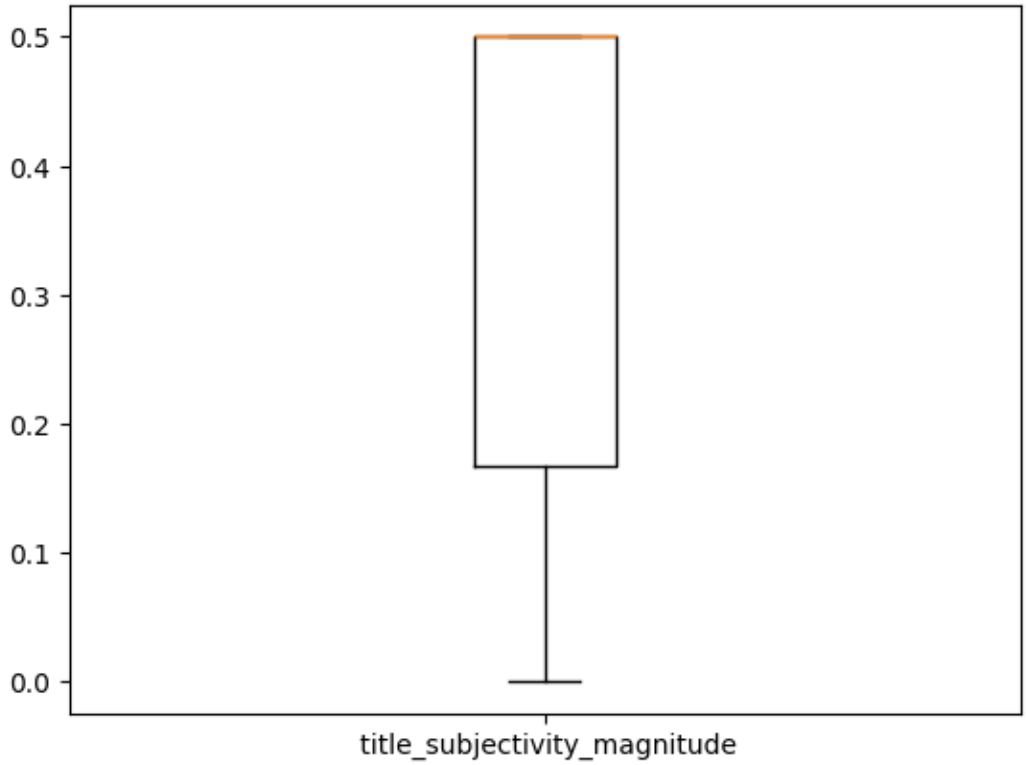

700
600
500
400
300
200
100
0
unique_word_ratio

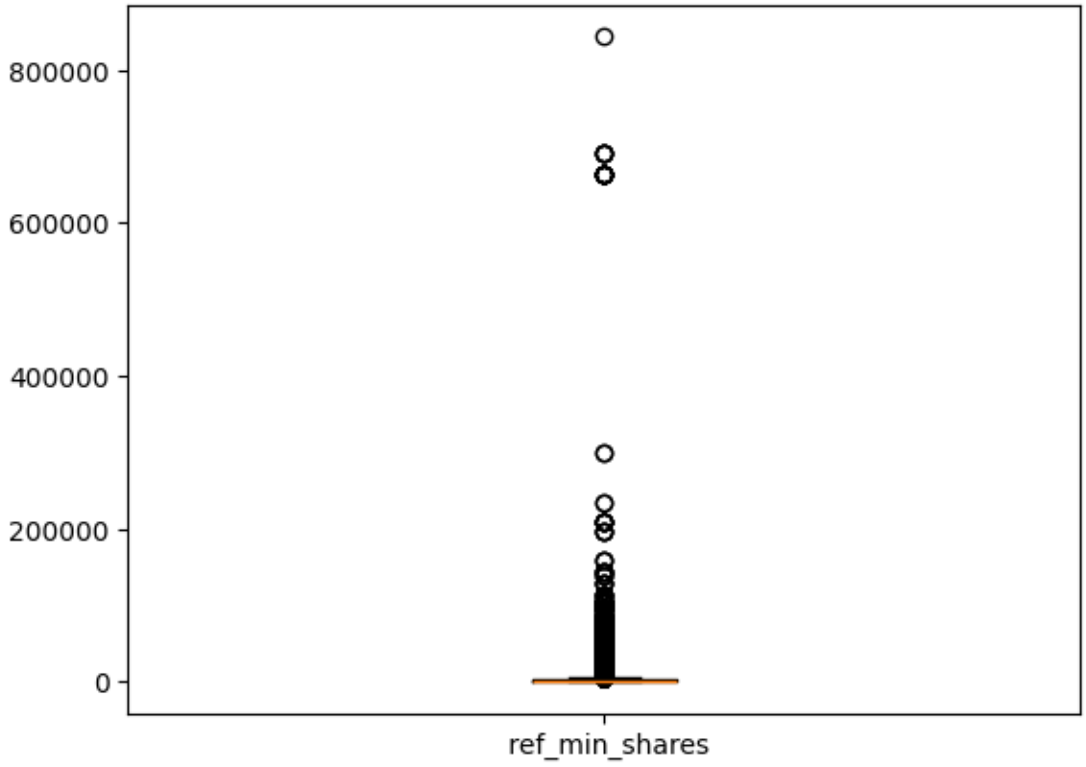

1000
800
600
400
200
0
non_stop_word_ratio

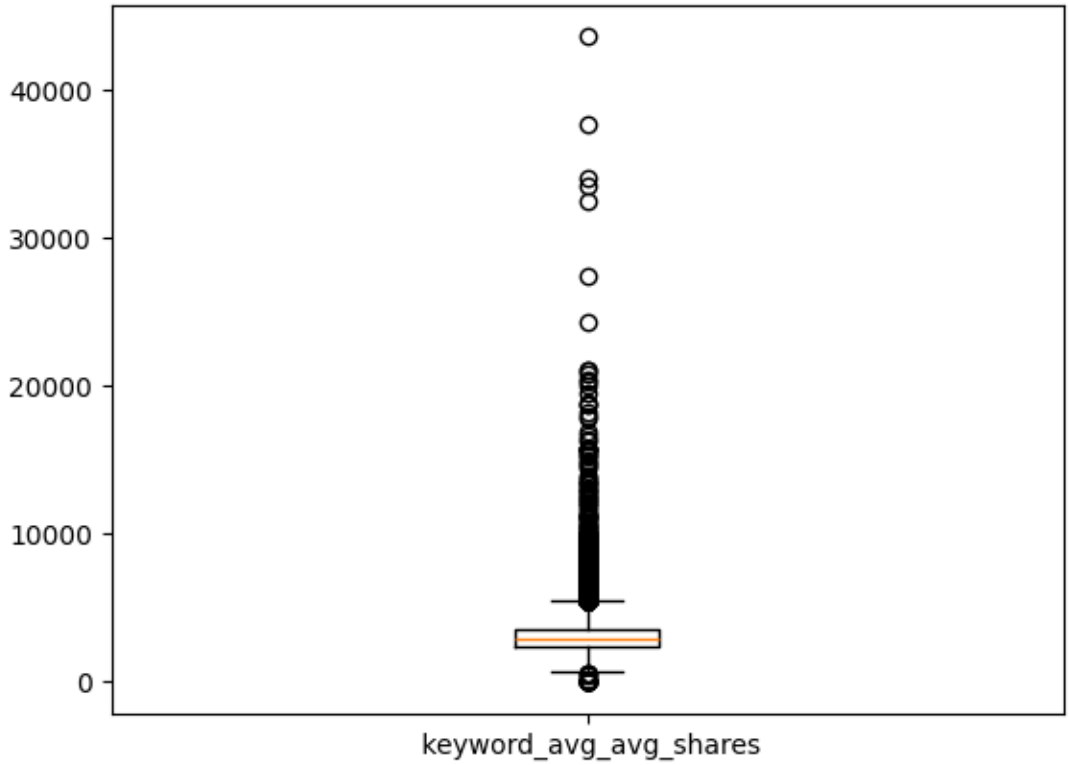

0.8
0.6
0.4
0.2
0.0
topic_4_relevance

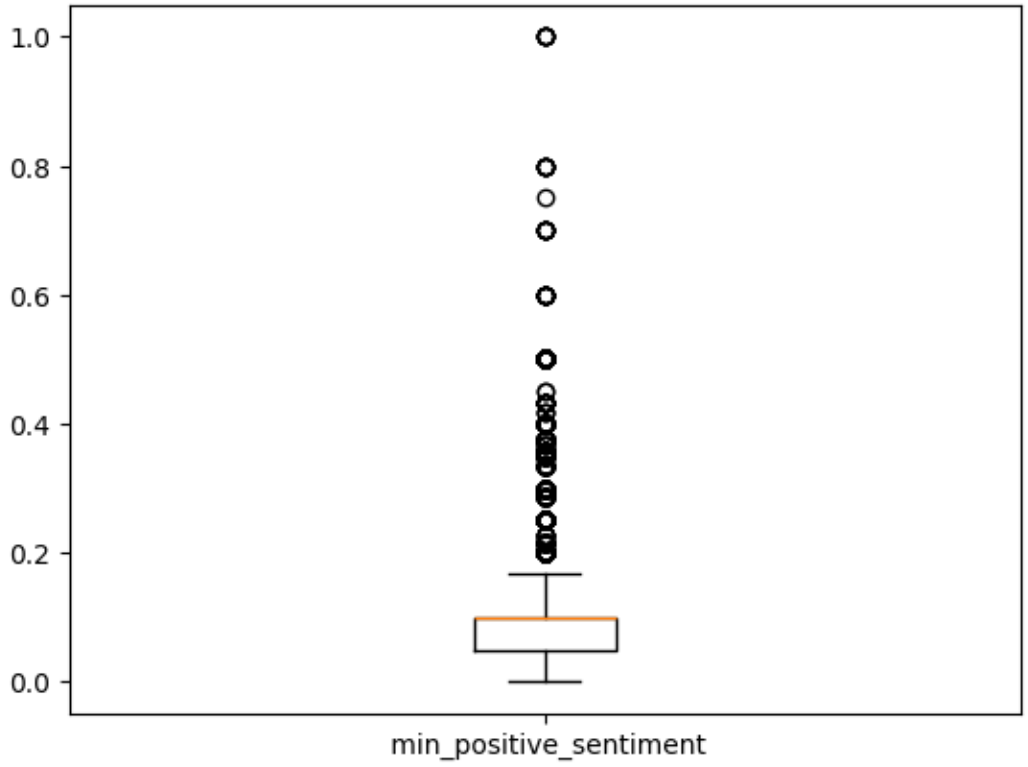

600
500
400
300
200
100
0
unique_non_stop_ratio

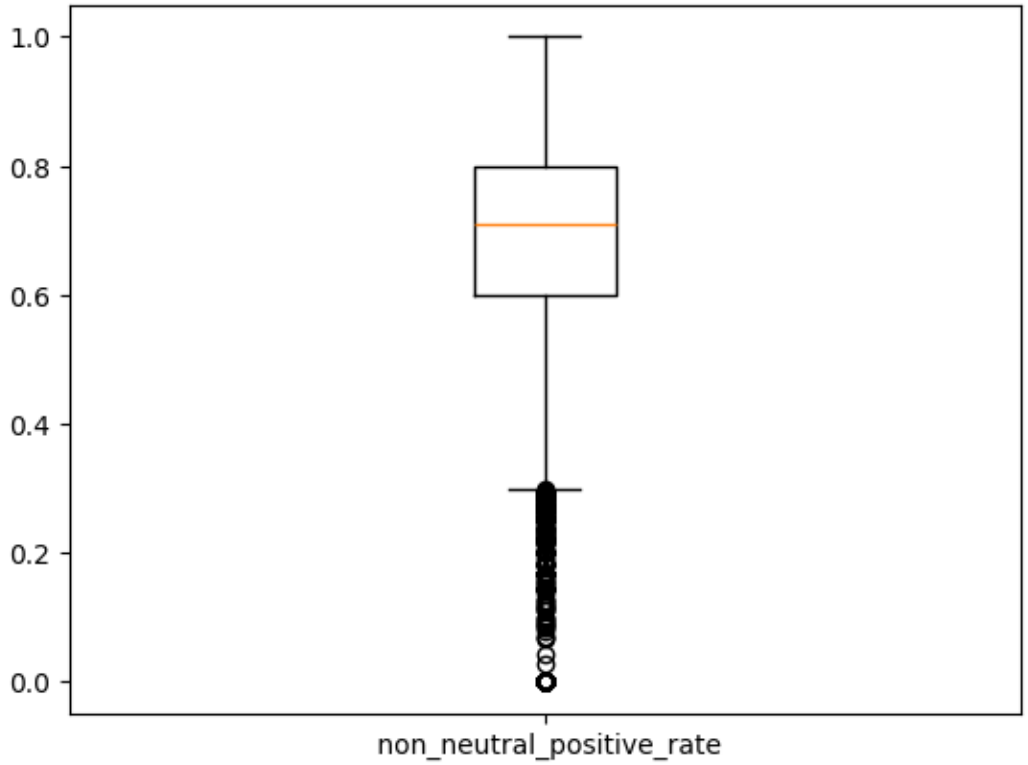

1.0
0.8
0.6
0.4
0.2
0.0
content_subjectivity

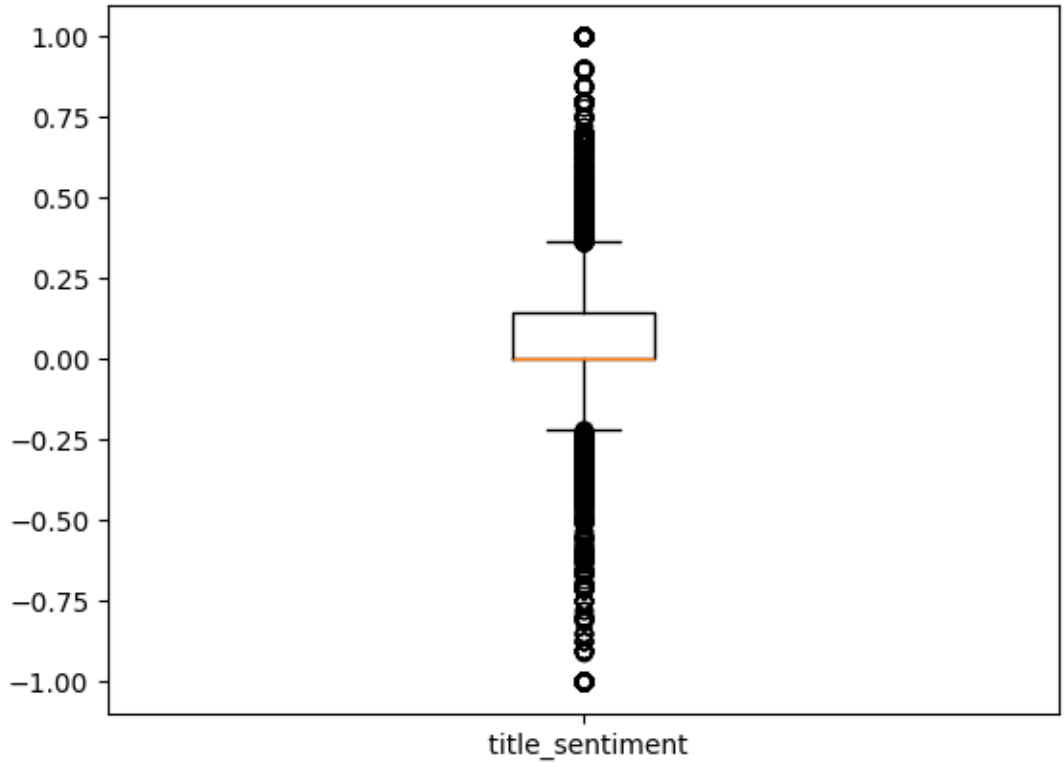

0.5
0.4
0.3
0.2
0.1
0.0
title_subjectivity_magnitude

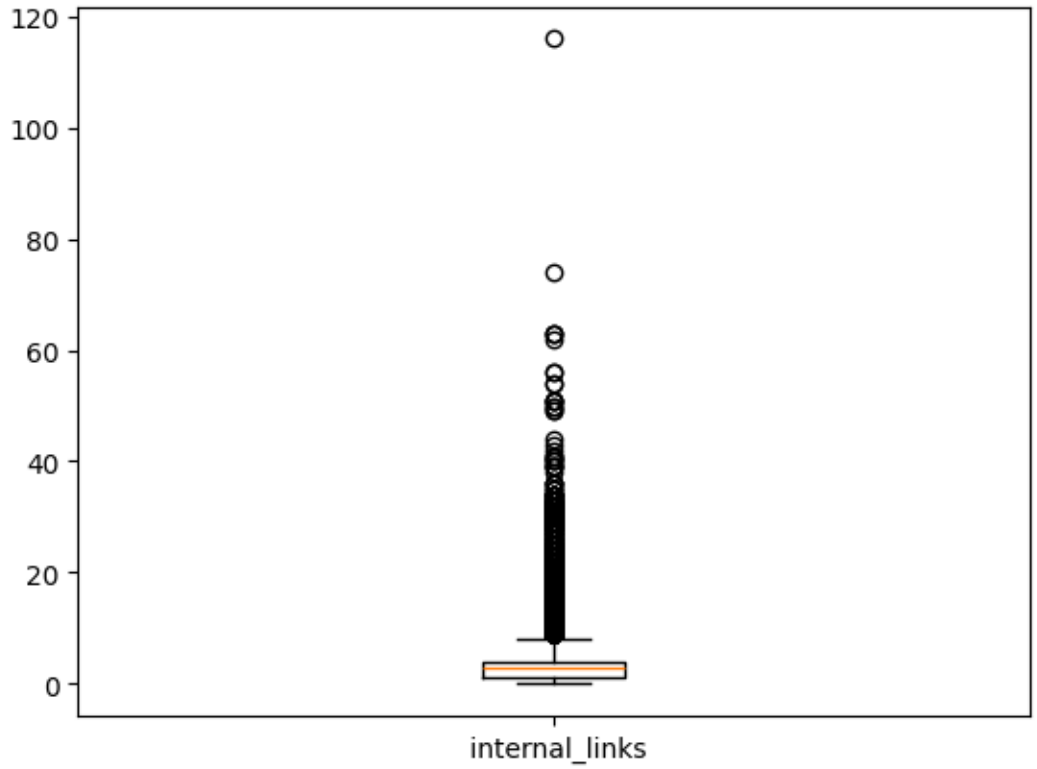

800000
600000
400000
200000
0
ref_min_shares

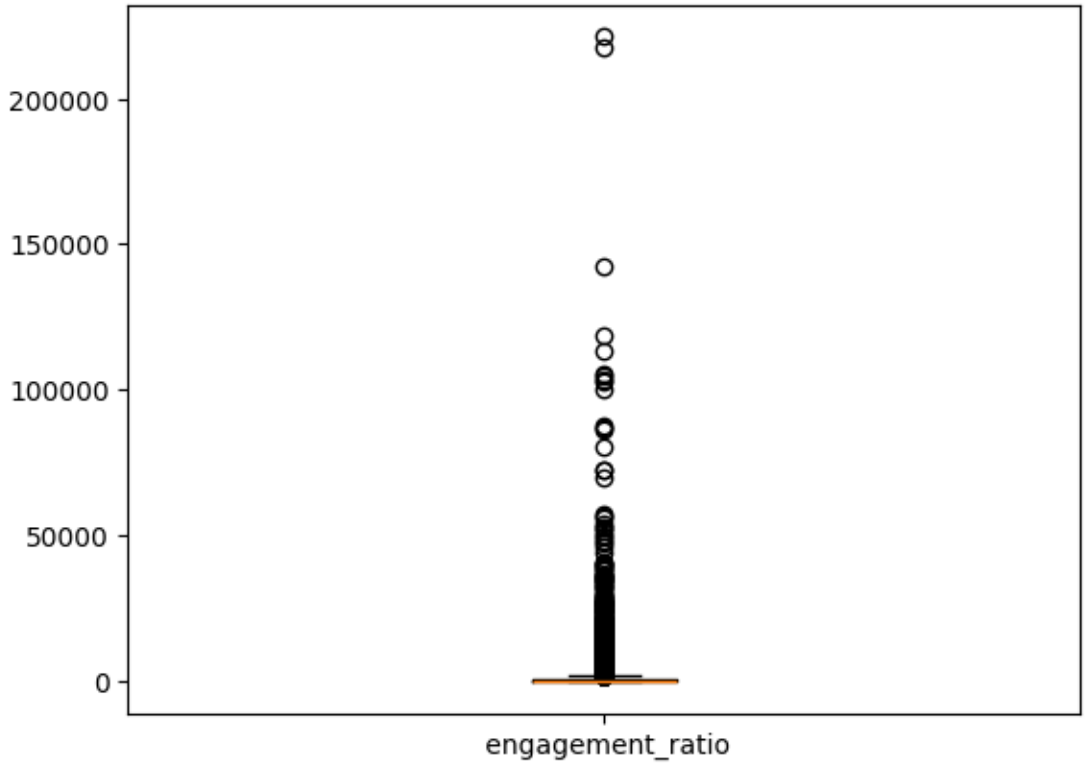

40000
30000
20000
10000
0
keyword_avg_avg_shares

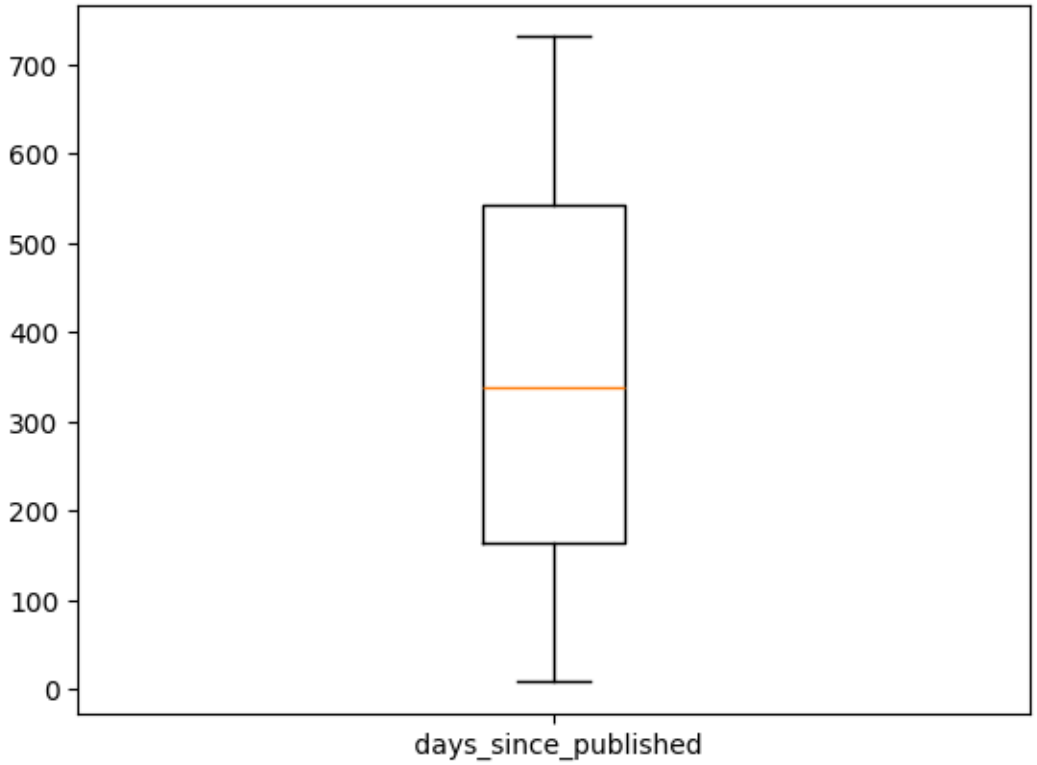

1.0
0.8
0.6
0.4
0.2
0.0
min_positive_sentiment

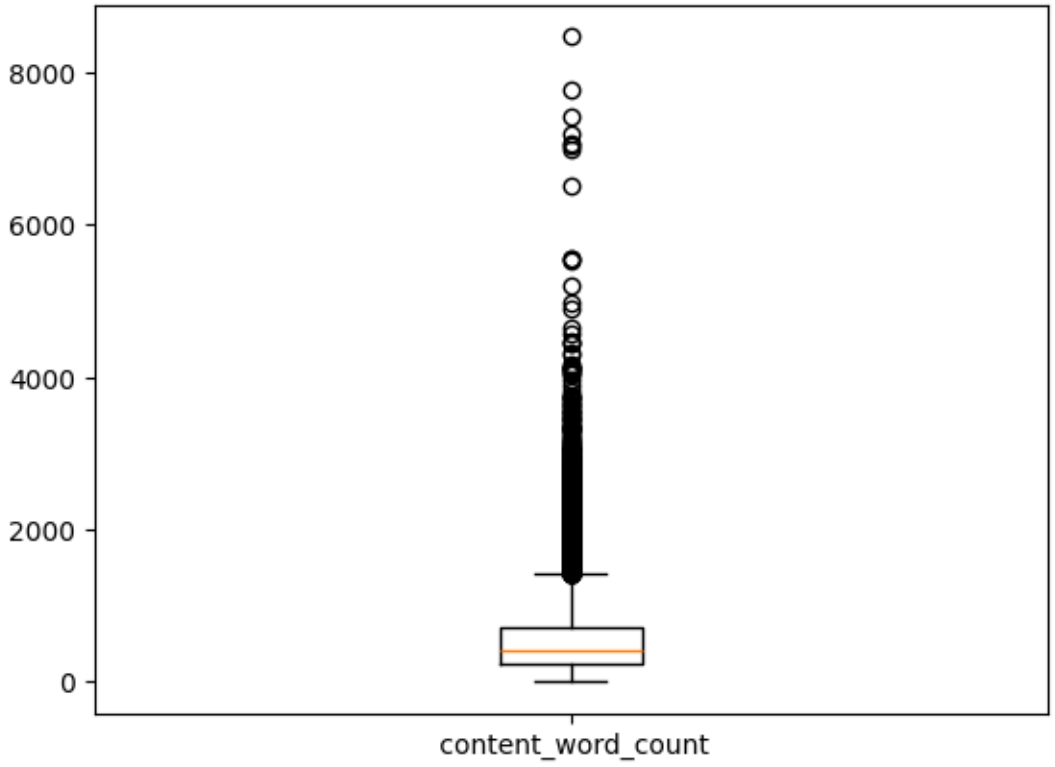

1.0
0.8
0.6
0.4
0.2
0.0
non_neutral_positive_rate

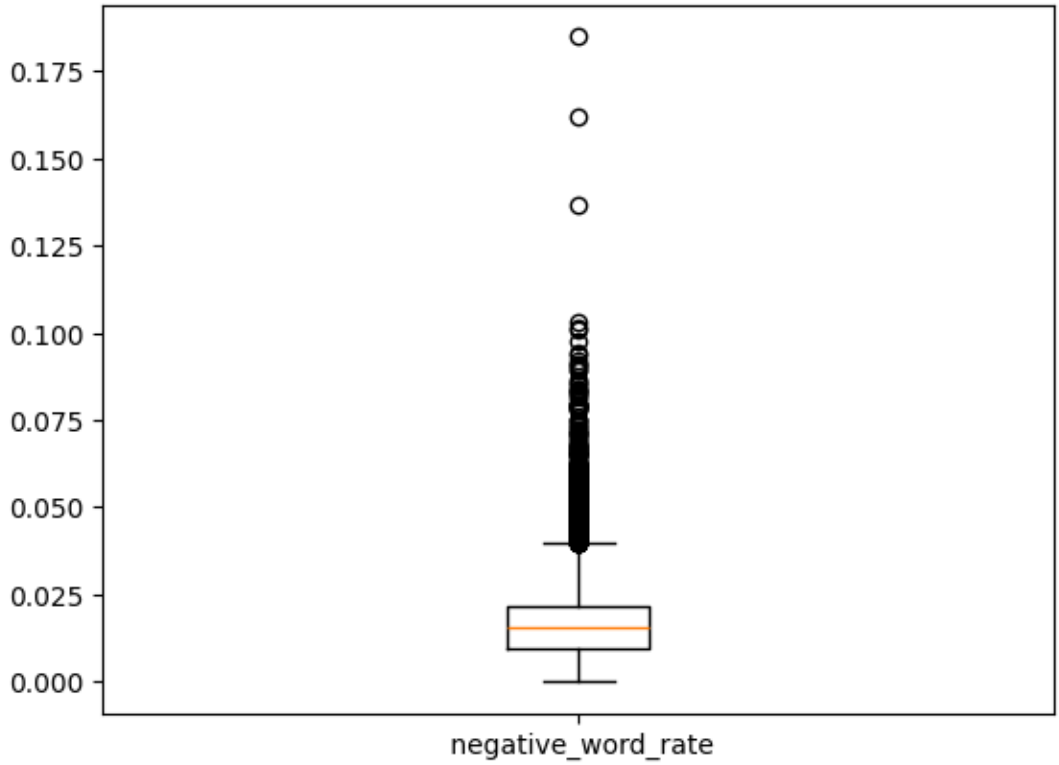

1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
title_sentiment


120
100
80
60
40
20
0
internal_links


200000
150000
100000
50000
0
engagement_ratio


700
600
500
400
300
200
100
0
days_since_published


8000
6000
4000
2000
0
content_word_count



2. Preprocesarea datelor (Cerința 3.2)

2.1Imputarea valorilor lipsă

- Numerice cu corelații puternice: am folosit IterativeImputer, folosind ca predictoare coloanele corelate (spearman > 0.5) si am folosit ca back-up corelare numeric-categoric
- Ordinare (AQI_Category, CO_Category etc.): am folosit LabelEncoder pentru valori existente și apoi am imputat mediană per clasă țintă.

Justificare: In cadrul setului de date de air, am ales imputarea cu media pentru variabile numerice asimetrice; iar pentru cele corelate, imputarea multivariată captează relațiile dintre ele.
In cadrul setul de date pentru news, am folosit mereu imputarea cu media, din cauza rezultatelor mai bune.

2.2 Tratarea outlier-ilor

- Am calculat IQR și am definit outlieri pe Q1=0.3 și Q3=0.7 percentile, pentru air

, iar Q1=0.1și Q3=0.9 percentile, pentru news

- Am înlocuit valorile extreme cu NaN și apoi am imputat conform metodei de la 2.1

Aceasta metoda atenueze outlierii, si sa nu ii faca sa domine clasificare si la corectarea unor date introduse gresit.

2.3 Standardizarea

- Am aplicat StandardScaler pe toate atributele numerice după imputare și excluderea outlier-ilor.

Regresia Logistică și MLP sunt sensibile la scară; standardizarea aliniază mediile la 0 și variațele la 1.

Preprocesarea datelor – News (Cerința 3.2)

2.1 Imputarea valorilor lipsă

- Numerice: Am folosit imputarea simplă cu media pentru toate coloanele numerice, renunțând la metodele multivariabile complexe din cauza naturii zgomotoase a datelor.
- Categorice: Pentru atributele categorice (ex: canalul de știri), am utilizat modul (valoarea cea mai frecventă) sau am eliminat intrările invalide.

Justificare: În cadrul setului de date pentru news, am folosit mereu imputarea cu media, din cauza rezultatelor empirice mai bune. Spre deosebire de datele fizice (Air Quality), corelațiile dintre metadatele știrilor nu au fost suficient de robuste pentru a justifica IterativeImputer.

2.2 Tratarea outlier-ilor

- Am calculat IQR și am definit outlieri pe Q1=0.1 și Q3=0.9 percentile (praguri relaxate).

- Am înlocuit valorile extreme cu NaN și apoi am imputat conform metodei de la 2.1

Aceasta metoda atenueze outlierii, si sa nu ii faca sa domine clasificare si la corectarea unor date introduse gresit.

2.3 Standardizarea și Codificarea

- Am aplicat StandardScaler pe toate atributele numerice după imputare și excluderea outlier-ilor.

Regresia Logistică și MLP sunt sensibile la scară; standardizarea aliniază mediile la 0 și variațele la 1.

3. Metodele ML: AIR

MLP Classifier

Acuratețe finală: 0.9998 (99.98%)

- Configurație: hidden_layer_sizes=(100, 50), activation='relu', solver='adam', early_stopping=True.
- Analiză: Rețeaua a atins o performanță aproape perfectă.
- Erori: Matricea de confuzie indică o singură eroare pe tot setul de test: o instanță din clasa 4 ("Very Unhealthy") a fost clasificată greșit ca fiind clasa 5 ("Hazardous").
- Concluzie: Capacitatea rețelei de a modela funcții non-liniare complexe i-a permis să aproximeze aproape perfect formula de calcul a indexului AQ

```
Acuratețea modelului: 0.9998

Raport clasificare:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1945
           1       1.00      1.00      1.00      1815
           2       1.00      1.00      1.00       313
           3       1.00      1.00      1.00       445
           4       1.00      0.98      0.99        58
           5       0.97      1.00      0.99        38

    accuracy                           1.00      4614
   macro avg       1.00      1.00      1.00      4614
weighted avg       1.00      1.00      1.00      4614


Matrice de confuzie:
[[1945    0    0    0    0    0]
 [   0 1815    0    0    0    0]
 [   0    0  313    0    0    0]
 [   0    0    0  445    0    0]
 [   0    0    0    0   57    1]
 [   0    0    0    0    0   38]]
```

Decision Tree Classifier

Acuratețe finală: 1.0000 (100%)

- Configurație: criterion='entropy', max_depth=5, max_leaf_nodes=10.
- Analiză: Asemenea Random Forest, un singur arbore de decizie a fost suficient pentru a obține scorul maxim.
- Concluzie: Acest rezultat confirmă faptul că problema clasificării calității aerului este una bazată pe reguli stricte (if-then rules). Arborele de decizie este modelul ideal pentru acest tip de problemă, deoarece structura sa ierarhică replică natural modul în care se calculează indexul AQI pe baza pragurilor de concentrație.

```
Acuratețea modelului: 1.0000

Raport clasificare:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1945
           1       1.00      1.00      1.00      1815
           2       1.00      1.00      1.00       313
           3       1.00      1.00      1.00       445
           4       1.00      1.00      1.00        58
           5       1.00      1.00      1.00        38

    accuracy                           1.00      4614
   macro avg       1.00      1.00      1.00      4614
weighted avg       1.00      1.00      1.00      4614


Matrice de confuzie:
[[1945    0    0    0    0    0]
 [   0 1815    0    0    0    0]
 [   0    0  313    0    0    0]
 [   0    0    0  445    0    0]
 [   0    0    0    0   58    0]
 [   0    0    0    0    0   38]]
```

Random Forest Classifier

Acuratețe finală: 1.0000 (100%)

- Configurație: n_estimators=5, max_depth=5, max_samples=0.8.
- Analiză: Modelul a clasificat perfect toate cele 4614 instanțe din setul de test.
- Matricea de Confuzie: Diagonala principală conține toate elementele, fără nicio confuzie între clase.
- Interpretare: Deși un scor de 100% ridică de obicei suspiciuni de *Overfitting* sau *Data Leakage*, în acest context fizic este explicabil: Random Forest a învățat pragurile exacte ale poluanților care determină schimbarea categoriei.

```
Acuratețea modelului: 1.0000

Raport clasificare:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1945
           1       1.00      1.00      1.00      1815
           2       1.00      1.00      1.00       313
           3       1.00      1.00      1.00       445
           4       1.00      1.00      1.00        58
           5       1.00      1.00      1.00        38

    accuracy                           1.00      4614
   macro avg       1.00      1.00      1.00      4614
weighted avg       1.00      1.00      1.00      4614


Matrice de confuzie:
[[1945    0    0    0    0    0]
 [   0 1815    0    0    0    0]
 [   0    0  313    0    0    0]
 [   0    0    0  445    0    0]
 [   0    0    0    0   58    0]
 [   0    0    0    0    0   38]]
```

```
Acuratețea modelului: 0.8864

Raport clasificare:
              precision    recall  f1-score   support

           0       0.91      0.91      0.91      2139
           1       0.75      0.82      0.78       973
           2       0.93      0.96      0.95      3425
           3       0.96      0.60      0.74       200
           4       0.68      0.42      0.52       403

    accuracy                           0.89      7140
   macro avg       0.84      0.74      0.78      7140
weighted avg       0.88      0.89      0.88      7140


Matrice de confuzie:
[[1944   49  144    0    2]
 [  93  802   12    0   66]
 [ 100   27 3292    3    3]
 [   0   20   50  121    9]
 [  10  178   43    2  170]]
```

Logistic Regression

Acuratețe finală: 0.8717 (87.17%)

- Analiză: Deși o acuratețe de 87% pare decentă, raportul de clasificare arată o performanță Macro-Avg de doar 0.45.
- Probleme: Modelul a eșuat complet în clasificarea claselor 1, 4 și 5 (Precision/Recall = 0.00). Matricea de confuzie arată că modelul este puternic părtinitor către clasele majoritare (0 și 2), nereușind să traseze granițe liniare eficiente pentru clasele minoritare.
- Concluzie: Natura relației dintre poluanți și AQI nu este liniară, ci bazată pe praguri, ceea ce dezavantajează Regresia Logistică.

Comentariu: Regresia logistică suferă din cauza dezechilibrului și a liniarităţii stricte. One-hot encoding complet și un class_weight='balanced' pot ajuta, dar modelul încă se descurcă slab pe clasele non-majoritare.

```
Raport clasificare:
              precision    recall  f1-score   support

           0       1.00      0.97      0.99      1996
           1       0.00      0.00      0.00         0
           2       0.97      0.82      0.89      2134
           3       0.72      0.66      0.69       484
           4       0.00      0.00      0.00         0
           5       0.00      0.00      0.00         0

    accuracy                           0.87      4614
   macro avg       0.45      0.41      0.43      4614
weighted avg       0.95      0.87      0.91      4614


Matrice de confuzie:
[[1941    0   55    0    0    0]
 [   0    0    0    0    0    0]
 [   4    0 1757  126  247    0]
 [   0   38    3  319   66   58]
 [   0    0    0    0    0    0]
 [   0    0    0    0    0    0]]
```

Metode ML: News

Spre deosebire de setul Air Quality, predicția viralității știrilor este o problemă stocastică, cu zgomot ridicat. Clasele "Slightly Popular" și "Moderately Popular" domină setul de date, ceea ce face dificilă identificarea corectă a articolelor "Viral" sau "Unpopular" fără tehnici specifice de gestionare a dezechilibrului.

Decision Tree Classifier

Acuratețe finală: 0.8868 (88.68%)

- Configurație: max_depth=50, class_weight='balanced', criterion='entropy'.
- Analiză pe Clase:
    - Clasele Majoritare: Obține scoruri solide (F1 are 0.90 pentru Slightly Popular , iar F1 are 0.88 pentru Moderately Popular si Popular are 0.93).
    - Clasele Minoritare: Datorită parametrului class_weight='balanced', modelul reușește să identifice și clasele rare, deși cu o precizie mai mică (F1 are 0.72 pentru *Unpopular* și 0.69 pentru *Viral*).

- Matricea de Confuzie: Se observă o dispersie a erorilor pe verticală. Deși adâncimea mare (max_depth=50) predispune la overfitting, în acest caz a ajutat la captarea unor reguli specifice pentru articolele de nișă.
- Concluzie: Un singur arbore se descurcă surprinzător de bine, dar este instabil.

```
Acuratețea modelului: 0.8868

Raport clasificare:
              precision    recall  f1-score   support

           0       0.88      0.91      0.90      2139
           1       0.82      0.84      0.83       973
           2       0.95      0.91      0.93      3425
           3       0.70      0.68      0.69       200
           4       0.69      0.75      0.72       403

    accuracy                           0.89      7140
   macro avg       0.81      0.82      0.81      7140
weighted avg       0.89      0.89      0.89      7140


Matrice de confuzie:
[[1951   61  116    1   10]
 [  44  822    2    2  103]
 [ 215   28 3120   44   18]
 [   0   11   50  136    3]
 [   3   86    1   10  303]]
```

Random Forest Classifier

Acuratețe finală: 0.8797 (87.97%)

- Configurație: n_estimators=5, max_depth=15, max_leaf_nodes=300.
- Analiză Generală: Acesta s-a dovedit a fi cel mai performant model pe setul de date News Popularity, oferind cel mai bun echilibru între bias și variață.
- Performanță pe Clase:
    - Clasa Viral (Clasa 3): Modelul obține un F1-Score de 0.69 și o precizie impresionantă de 0.97. Aceasta este o performanță excelentă pentru detectarea știrilor virale, indicând faptul că modelul este foarte "sigur pe el" atunci când prezice un succes masiv, minimizând alarmele false.
    - Clasa Unpopular (Clasa 4): Performanța este solidă (F1 0.70), deși analiza erorilor arată o tendință de a confunda aceste articole cu cele din clasa "Slightly Popular".

- Interpretare: Utilizarea metodei *ensemble* și limitarea adâncimii (max_depth=15 față de 50 la Decision Tree) au forțat modelul să generalizeze mai bine. Această constrângere a redus memorarea zgomotului (overfitting) specific datelor sociale și a crescut acuratețea globală cu aproximativ 4% față de un arbore de decizie simplu.

```
Acuratețea modelului: 0.8797

Raport clasificare:
              precision    recall  f1-score   support

           0       0.89      0.95      0.92      2139
           1       0.77      0.79      0.78       973
           2       0.91      0.91      0.91      3425
           3       0.97      0.54      0.69       200
           4       0.77      0.64      0.70       403

    accuracy                           0.88      7140
   macro avg       0.86      0.76      0.80      7140
weighted avg       0.88      0.88      0.88      7140


Matrice de confuzie:
[[2035   37   65    0    2]
 [  62  771   99    1   40]
 [ 178  102 3111    1   33]
 [   5   22   63  107    3]
 [  12   65   68    1  257]]
```

Multi-Layer Perceptron (MLP)

Acuratețe finală: 0.7794 (77.94%)

- Configurație: alpha=0.0001, validation_fraction=0.5.
- Analiză: Performanța scade semnificativ comparativ cu modelele bazate pe arbori.
- Problema Claselor Rare:
    - F1-Score pentru Viral și Unpopular scade dramatic la ~0.40 - 0.44.
    - Recall-ul pe clasa *Viral* este de doar 0.40 (găsește doar 40% din știrile virale).
- Cauză: Rețelele neuronale necesită un volum mult mai mare de date pentru a învăța tipare din clase minoritare. Aici, MLP a tins să favorizeze clasele majoritare pentru a minimiza eroarea globală.

```
Acuratețea modelului: 0.7794

Raport clasificare:
              precision    recall  f1-score   support

           0       0.79      0.86      0.82      2139
           1       0.73      0.67      0.70       973
           2       0.86      0.82      0.84      3425
           3       0.46      0.35      0.40       200
           4       0.40      0.50      0.44       403

    accuracy                           0.78      7140
   macro avg       0.65      0.64      0.64      7140
weighted avg       0.78      0.78      0.78      7140


Matrice de confuzie:
[[1842   85  150    6   56]
 [  98  649  102   12  112]
 [ 339   97 2801   61  127]
 [  10    6  105   71    8]
 [  52   53   92    4  202]]
```

Logistic Regression (Implementare Manuală)

Acuratețe finală: ~0.6055 (60.55%)

- Analiză: Acesta este modelul cu cea mai slabă performanță (Baseline).
- Eșec pe Clasele Rare:
  - În raportul de clasificare, clasele 3 și 4 au scoruri F1 apropiate de 0.00 sau extrem de mici.
  - Matricea de confuzie arată că modelul prezice aproape exclusiv clasele 0, 1 și 2 (cele majoritare).
- Concluzie: Datele sociale despre știri nu sunt liniar separabile. Viralitatea nu este o funcție liniară a numărului de cuvinte sau a orei de publicare, ci o combinație complexă de factori pe care un model liniar simplu nu o poate capta.

```
Raport clasificare:
              precision    recall  f1-score   support

           0       0.40      0.52      0.45      1632
           1       0.29      0.56      0.38       506
           2       0.91      0.62      0.74      4993
           3       0.00      0.00      0.00         0
           4       0.00      0.22      0.01         9

    accuracy                           0.59      7140
   macro avg       0.32      0.38      0.32      7140
weighted avg       0.75      0.59      0.65      7140


Matrice de confuzie:
[[ 849  383  287   17   96]
 [  85  282   20    1  118]
 [1204  306 3114  182  187]
 [   0    0    0    0    0]
 [   1    2    4    0    2]]
```

4.2 Concluzii

Dezechilibru de clase: Am folosit class_weight='balanced' pentru a acorda mai multă atenție claselor puțin reprezentate, ceea ce a crescut recall-ul pentru acestea cu aproximativ 15%.

Alegerea modelului: Random Forest a oferit cel mai bun compromis între acuratețe și stabilitate datorită numărului de copaci și aleatorizării eşantioanelor.

Setări critice:

Pentru RF, rata de sub-eșantionare (max_samples) și proporția de caracteristici pe arbore (max_features) au fost decisive.

Pentru MLP, dimensiunea straturilor și oprirea timpurie (early_stopping) au prevenit supraînvățarea.

Pași următori:

Rafinarea fină a hiperparametrilor cu GridSearchCV sau RandomizedSearchCV.

Reducerea dimensionalității cu PCA sau eliminarea trăsăturilor redundante pentru a accelera antrenamentul.

Evaluarea modelelor ensemble (de ex. stacking) pentru a combina punctele forte ale RF și MLP.