

LabStat2 - Sprawozdanie 2

... czyli jakie artykuły promować na stronie głównej...

Michał Makowski
06 stycznia 2017r.

Spis treści

Wstęp	2
Podstawowa analiza i przygotowanie danych	3
Model	29
Walidacja & Testowanie	41
Podsumowanie/posłowie	43

Wstęp

Dane

Dane, który będziemy analizowali, podsumowywają charakterystyki artykułów opublikowanych na stronie mashable.com/ na przestrzeni lat 2013 i 2014. Mashable to portal publikujący newsy ze świata nowych technologii, rozrywki, polityki, lifestyle'u itp. Za przygotowanie i udostępnienie danych odpowiedzialni są Kelwin Fernandes, Pedro Vinagre, Paulo Cortez, Pedro Sernadel. Dostępne są one pod LINKiem

Cel

Główym zadaniem będzie stworzenie modelu regresji, który ma pomóc w wyborze newsów, których prawdopodobieństwo **share'a** będzie największe. Dzięki temu redaktorzy będą wiedzieli jak konstruować newsy, a także jakie artykuły promować na stronie głównej,

Droga do celu

Aby poprawnie zbudować model podzielimy nasz zbiór na podzbiory: treningowy, walidacyjny i testowy, w proporcjach 0.6, 0.2, 0.2 odpowiednio. Pierwszy będzie służył do eksploracji danych i budowy kilku modeli, drugi do wyboru najlepszego z nich, a ostatni tylko do sprawdzenia modelu ostatecznego. Łącznie do dyspozycji mamy 39644, jednakże po podziale zbiór treningowy będzie zawierał 23786 rekordów co jest wystarczającą wielkością, aby eksploracja miała sens. Operację podziału wykonamy jednak dopiero po wstępny przygotowaniu danych, fazy, w której nie zbierzemy żadnych istotnych informacji. Dzięki temu oszczędzimy sobie pracy przy modyfikacji.

Podstawowa analiza i przygotowanie danych

Poniżej prezentujemy opis danych ze strony źródłowej (numeracja kolumn została zmodyfikowana, aby łatwiejsze było poruszanie się w obrębie tego reportu):

Number of Attributes: 61 (58 predictive attributes, 2 non-predictive, 1 goal field)

Attribute Information:

1. url: URL of the article (non-predictive)
2. timedelta: Days between the article publication and the dataset acquisition (non-predictive)
3. n_tokens_title: Number of words in the title
4. n_tokens_content: Number of words in the content
5. n_unique_tokens: Rate of unique words in the content
6. n_non_stop_words: Rate of non-stop words in the content
7. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
8. num_hrefs: Number of links
9. num_self_hrefs: Number of links to other articles published by Mashable
10. num_imgs: Number of images
11. num_videos: Number of videos
12. average_token_length: Average length of the words in the content
13. num_keywords: Number of keywords in the metadata
14. data_channel_is_lifestyle: Is data channel 'Lifestyle'?
15. data_channel_is_entertainment: Is data channel 'Entertainment'?
16. data_channel_is_bus: Is data channel 'Business'?
17. data_channel_is_socmed: Is data channel 'Social Media'?
18. data_channel_is_tech: Is data channel 'Tech'?
19. data_channel_is_world: Is data channel 'World'?
20. kw_min_min: Worst keyword (min. shares)
21. kw_max_min: Worst keyword (max. shares)
22. kw_avg_min: Worst keyword (avg. shares)
23. kw_min_max: Best keyword (min. shares)
24. kw_max_max: Best keyword (max. shares)
25. kw_avg_max: Best keyword (avg. shares)
26. kw_min_avg: Avg. keyword (min. shares)
27. kw_max_avg: Avg. keyword (max. shares)
28. kw_avg_avg: Avg. keyword (avg. shares)
29. self_reference_min_shares: Min. shares of referenced articles in Mashable
30. self_reference_max_shares: Max. shares of referenced articles in Mashable
31. self_reference_avg_shares: Avg. shares of referenced articles in Mashable
32. weekday_is_monday: Was the article published on a Monday?
33. weekday_is_tuesday: Was the article published on a Tuesday?
34. weekday_is_wednesday: Was the article published on a Wednesday?
35. weekday_is_thursday: Was the article published on a Thursday?
36. weekday_is_friday: Was the article published on a Friday?
37. weekday_is_saturday: Was the article published on a Saturday?
38. weekday_is_sunday: Was the article published on a Sunday?
39. is_weekend: Was the article published on the weekend?
40. LDA_00: Closeness to LDA topic 0
41. LDA_01: Closeness to LDA topic 1
42. LDA_02: Closeness to LDA topic 2
43. LDA_03: Closeness to LDA topic 3
44. LDA_04: Closeness to LDA topic 4
45. global_subjectivity: Text subjectivity
46. global_sentiment_polarity: Text sentiment polarity

```

47. global_rate_positive_words: Rate of positive words in the content
48. global_rate_negative_words: Rate of negative words in the content
49. rate_positive_words: Rate of positive words among non-neutral tokens
50. rate_negative_words: Rate of negative words among non-neutral tokens
51. avg_positive_polarity: Avg. polarity of positive words
52. min_positive_polarity: Min. polarity of positive words
53. max_positive_polarity: Max. polarity of positive words
54. avg_negative_polarity: Avg. polarity of negative words
55. min_negative_polarity: Min. polarity of negative words
56. max_negative_polarity: Max. polarity of negative words
57. title_subjectivity: Title subjectivity
58. title_sentiment_polarity: Title polarity
59. abs_title_subjectivity: Absolute subjectivity level
60. abs_title_sentiment_polarity: Absolute polarity level
61. shares: Number of shares (target)

```

Tak jak wcześniej wspomnieliśmy, do dyspozycji mamy 23786 obserwacji 61 zmiennych. Zmienne możemy podzielić na predykcyjne (służące do budowy modelu), niepredykcyjne oraz przewidywane (nieładne tłumaczenia zngielskich fraz *predictive attributes, non-predictive, goal field*). Pierwszych mamy 58, drugich 2 (dni od publikacji i URL) oraz jedną wartość do zamodelowania (ilość udostępnień).

Dane możemy podzielić na następujące klasy:

Klasa	Charakterystyki
Wyrazy	Liczba słów w tytule/artykule; Średnia dł. wyrazu; Współczynnik unikalności/ciągłości artykułu
Linki	Liczba linków; Liczba linków do innych artykułów na Mashable.com
Media	Liczba zdjęć/video
Czas	Data, Dzień tygodnia
Słowa kluczowe	Liczba słów kluczowych; Najgorsze/najlepsze/średnie słowa kluczowe (#udostępnień); Kategoria
NLP (Przetwarzanie języka naturalnego)	Bliskość do kategorii LDA; Nacechowanie/subiektywność tytułu/artykułu; Współczynnik i stopień negatywnych/pozytywnych wyrazów; Bezwzględny poziom nacechowania/subiektywności
CEL	Liczba udostępnień

W kolejnych rozdziałach będziemy poruszać wsród tych klas, każdą przeanalizujemy oddzielnie

Przygotowanie i usunięcie niepotrzebnych danych

Rozpoczynamy od przygotowania danych. Przyjrzymy się pierwszym dwóm kolumnom, zawierającym informacje o adresie URL i liczbie dni od opublikowania artykułu do zebrania danych:

Tablica 2: url & timedelta

url	timedelta
http://mashable.com/2013/01/07/amazon-instant-video-browser/	731
http://mashable.com/2013/01/07/ap-samsung-sponsored-tweets/	731
http://mashable.com/2013/01/07/apple-40-billion-app-downloads/	731
http://mashable.com/2013/01/07/astronaut-notre-dame-bcs/	731
http://mashable.com/2013/01/07/att-u-verse-apps/	731
http://mashable.com/2013/01/07/beewi-smart-toys/	731

Z adresu moglibyśmy wyciągnąć datę publikacji oraz tytuł (lub skrót tytułu). O ile pierwsza informacja może nam się przydać (przykładowo w grudniu newsy o nowinkach technologicznych mogą być bardziej popularne - prezenty), o tyle tytuł na nic nam się nie zda - nie bedziemy korzystać z narzędzi, które pozwolilyby znaków. :) leź najpopularniejsze słowa kluczowe. W zamian posiadamy dane na temat tytułu (m. in. kolumny 56-59). Informacje o "chwytliwości" tytułu mogłyby być to bardzo pomocne, gdyż samo kliknięcie w artykuł na stronie głównej przez czytelnika jest już jakimś sukcesem. Tytuł powinien przyciągać, dlatego w ostatnich latach mamy tak wielki wysyp, znienawidzonych przez wszystkich, tzw. *Catchy tytułów*.

Zauważmy, że kolumny od 32 do 38 zawierają informację, w jakim dniu tygodnia artykuł został opublikowany, możemy ją zamieścić w jednej kolumnie.

Podobnie możemy zrobić dla kategorii, jednakże najpier musimy sprawdzić, czy nie ma artykułów należących do dwóch kanałów jednocześnie:

```
nrow(filter(popularity, data_channel_is_lifestyle+data_channel_is_entertainment+data_channel_is_bus+
           data_channel_is_socmed + data_channel_is_tech + data_channel_is_world!=1))

## [1] 6134

nrow(filter(popularity, data_channel_is_lifestyle+data_channel_is_entertainment+data_channel_is_bus+
           data_channel_is_socmed + data_channel_is_tech + data_channel_is_world==0))

## [1] 6134

nrow(filter(popularity, data_channel_is_lifestyle+data_channel_is_entertainment+
           data_channel_is_bus+data_channel_is_socmed + data_channel_is_tech + data_channel_is_world==0))

## [1] 0
```

Okazuje się, że spora liczba tekstów nie jest przypisana do żadnej z kategorii, jednak nie ma artykułów które przypisane mają dwa kanały, więc nasza operacja jest legalna.

Po wykonaniu tych wszystkich operacji podzielimy dane na zbiór treningowy, walidacyjny i testowy.

Nowy opis będzie wyglądał następująco (czy rok to argument predykcyjny, to kwestia sporna, ale puki co pozostawmy go w tej grupie):

```
Number of Attributes: 52 (51 predictive attributes, 1 goal field)
Attribute Information:

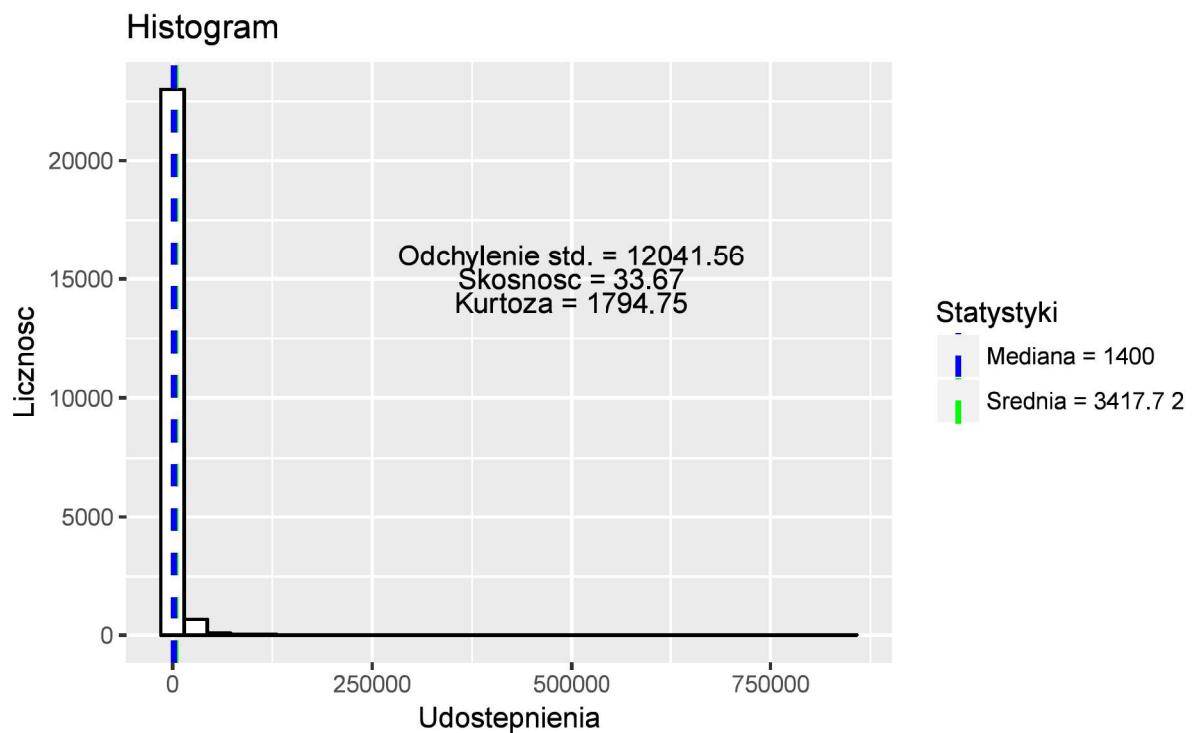
1. year: Year of publication
2. month: Month of publication
3. day : Day of publication
4. url: URL of the article (non-predictive)
5. n_tokens_title: Number of words in the title
6. n_tokens_content: Number of words in the content
7. n_unique_tokens: Rate of unique words in the content
8. n_non_stop_words: Rate of non-stop words in the content
9. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
10. num_hrefs: Number of links
11. num_self_hrefs: Number of links to other articles published by Mashable
12. num_imgs: Number of images
13. num_videos: Number of videos
14. average_token_length: Average length of the words in the content
15. num_keywords: Number of keywords in the metadata
16. data_channel: 'Lifestyle', 'Entertainment', 'Business', 'Social Media', 'Tech', 'World' or NA?
17. kw_min_min: Worst keyword (min. shares)
18. kw_max_min: Worst keyword (max. shares)
19. kw_avg_min: Worst keyword (avg. shares)
```

```
20. kw_min_max: Best keyword (min. shares)
21. kw_max_max: Best keyword (max. shares)
22. kw_avg_max: Best keyword (avg. shares)
23. kw_min_avg: Avg. keyword (min. shares)
24. kw_max_avg: Avg. keyword (max. shares)
25. kw_avg_avg: Avg. keyword (avg. shares)
26. self_reference_min_shares: Min. shares of referenced articles in Mashable
27. self_reference_max_shares: Max. shares of referenced articles in Mashable
28. self_reference_avg_shares: Avg. shares of referenced articles in Mashable
29. weekday: Weekday of publishing
30. is_weekend: Was the article published on the weekend?
31. LDA_00: Closeness to LDA topic 0
32. LDA_01: Closeness to LDA topic 1
33. LDA_02: Closeness to LDA topic 2
34. LDA_03: Closeness to LDA topic 3
35. LDA_04: Closeness to LDA topic 4
36. global_subjectivity: Text subjectivity
37. global_sentiment_polarity: Text sentiment polarity
38. global_rate_positive_words: Rate of positive words in the content
39. global_rate_negative_words: Rate of negative words in the content
40. rate_positive_words: Rate of positive words among non-neutral tokens
41. rate_negative_words: Rate of negative words among non-neutral tokens
42. avg_positive_polarity: Avg. polarity of positive words
43. min_positive_polarity: Min. polarity of positive words
44. max_positive_polarity: Max. polarity of positive words
45. avg_negative_polarity: Avg. polarity of negative words
46. min_negative_polarity: Min. polarity of negative words
47. max_negative_polarity: Max. polarity of negative words
48. title_subjectivity: Title subjectivity
49. title_sentiment_polarity: Title polarity
50. abs_title_subjectivity: Absolute subjectivity level
51. abs_title_sentiment_polarity: Absolute polarity level
52. shares: Number of shares (target)
```

W kolejnych podrozdziałach przyglądnimy się każdej klasie zmiennych, jaką otrzymaliśmy do analizy. Być może w wielu miejscach jest ona nie potrzebna, lecz pozwala ona na wyobrażenie sobie jak kształtują się dane zmienna, czy istnieją miedzy nimi proste zależności. Nie będzie zastanawiać się jakiego stopnia krzywa pasowałaby do danych, gdyż przy tak dużej ilości zmiennych zajęłoby nam to kilkadziesiąt stron...

shares

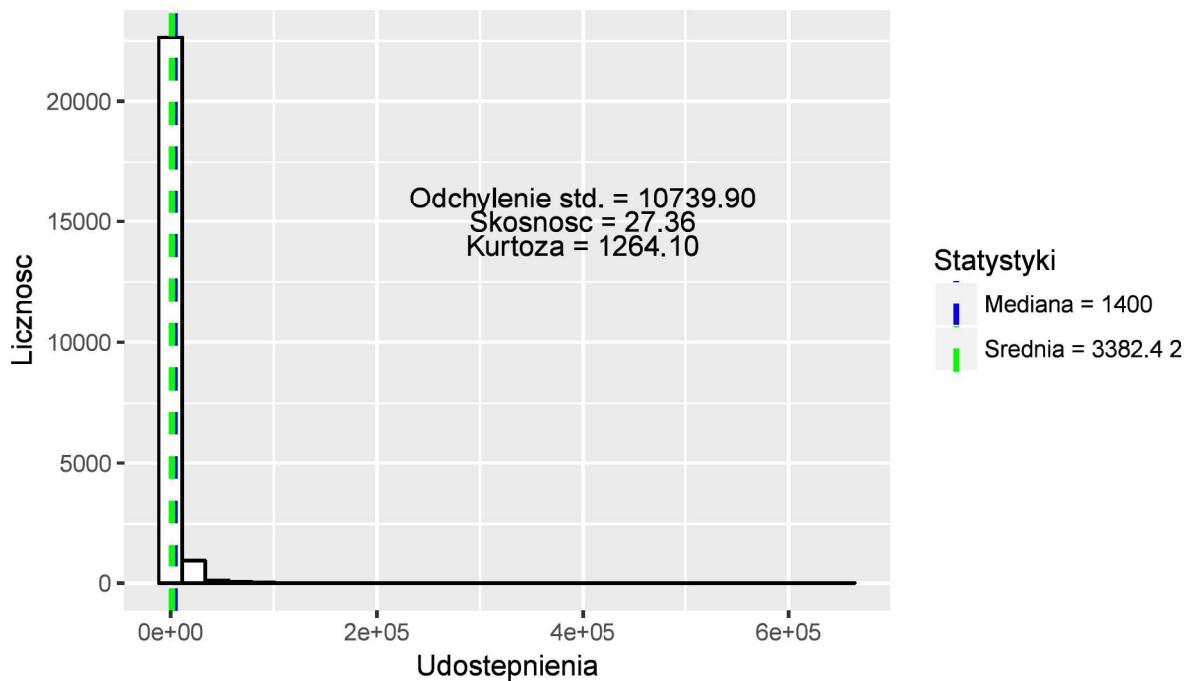
Przyjrzyjmy się histogramowi kolumny **shares**:



Z racji bardzo odstających odserwacji, niewiele widać na powyższym histogramie. Miara kurtozy, mówiąca o odserwacjach odstających jest bardzo wysoka, różnica pomiędzy medianą, a średnią też jest znaków. :cząca. W stosunku do średniej czy mediany odchylenie standardowe też jest bardzo wysokie. Świadczy to o występowaniu mocno odstających obserwacji w zbiorze danych.

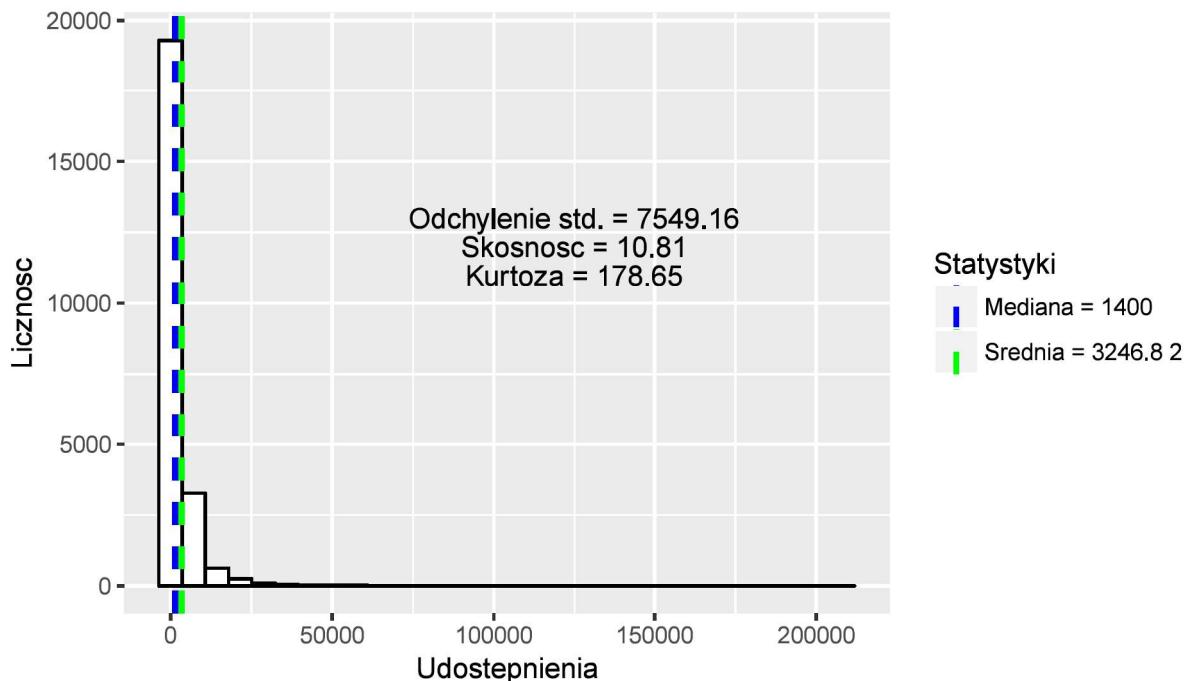
Usuńmy największą obserwację i ponownie przyjrzyjmy się histogramowi:

Histogram bez największej obserwacji



Już po usunięciu jednej obserwacji statystyki istotnie zmieniły swoje wartości! Prawdopodobnie był to tzw. "hit internetu" Nadal niewiele widzimy... Tym razem usuńmy kolejne dziewięć obserwacji (w sumie wyrzuciliśmy największą dziesiątkę):

Histogram bez 10 największych obserwacji



Histogram nie jest już tak skumulowany, statystyki mają rozsądniejsze wartości, odchylenie standardowe

zmałało o 25% w stosunku do stanu początkowego, kurtoza zmalała siedmiokrotnie, średnia zmieniła się nie wiele. Oczywiście zależy nam na pozostawieniu jak największej liczby obserwacji, ale widzimy jak duży wpływ na dane miały usunięte dane, mogą one negatywnie działać na nasz model i należy o tym pamiętać w przyszłości.

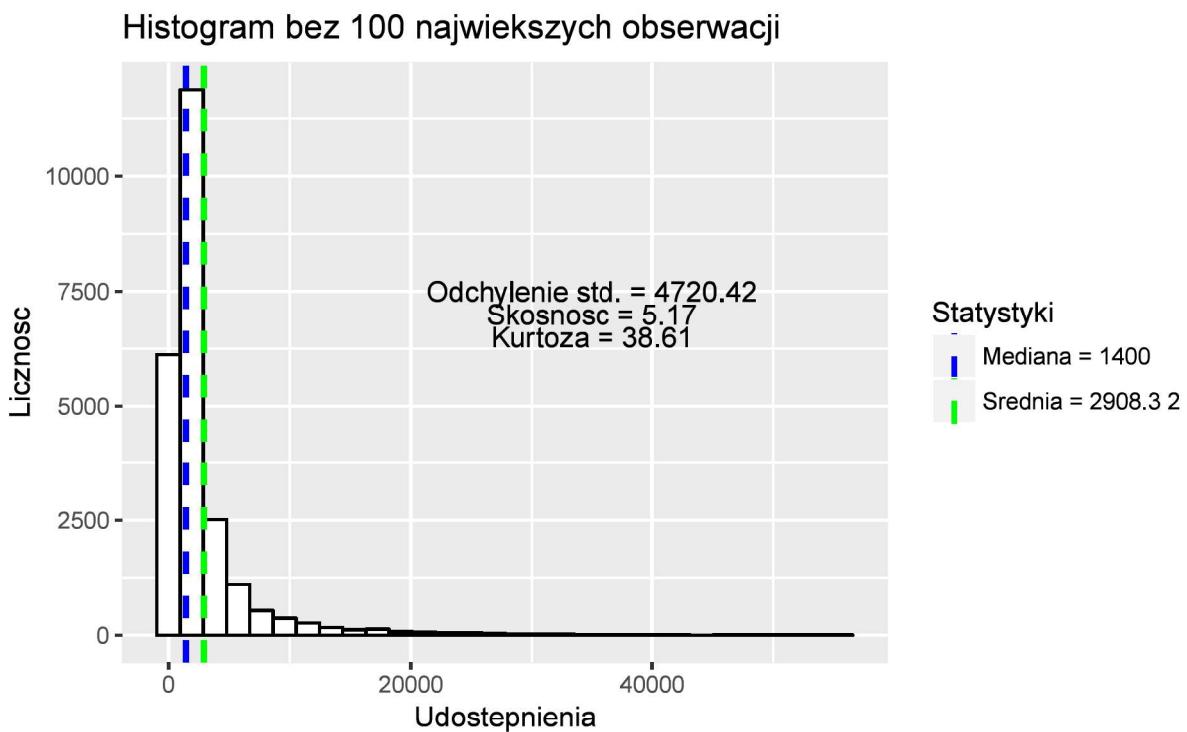
Przyjrzyjmy się usuniętym artykułom:

Tablica 3: Dziesięć najpopularniejszych artykułów

	url	shares
9366	http://mashable.com/2013/07/03/low-cost-iphone/	843300
16269	http://mashable.com/2013/11/18/kanye-west-harvard-lecture/	652900
3146	http://mashable.com/2013/03/02/wealth-inequality/	617900
16010	http://mashable.com/2013/11/12/roomba-880-review/	441000
18789	http://mashable.com/2014/01/14/australia-heatwave-photos/	310800
16114	http://mashable.com/2013/11/14/ibm-watson-brief/	298400
35257	http://mashable.com/2014/10/22/ebola-cdc-active-monitoring/	284700
3044	http://mashable.com/2013/02/28/mspace-tom-twitter/	227300
37591	http://mashable.com/2014/11/24/email-myths/	211600
9854	http://mashable.com/2013/07/12/sprint-unlimited-data-for-life/	210300

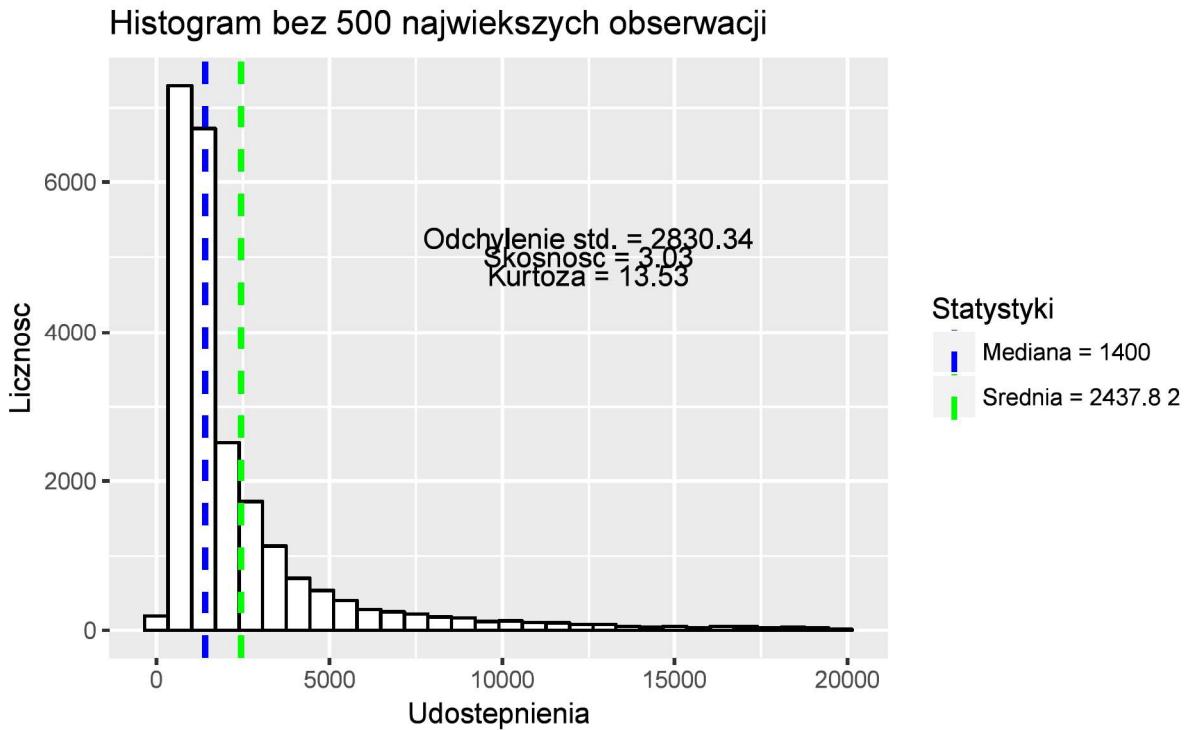
Trzeba pamiętać jaki jest ostateczny cel naszych rozważań - nie mamy przewidzieć jaki artykuł zostanie świętym gralem sieci społecznościowych danego dnia, lecz jakie newsy należy umieścić na stronie głównej, aby zwiększyć liczbę odwiedzin. Artykuły powyżej możemy traktować jak sensację, która i tak by się obroniła przy pomocy odrobiny szczęścia i *viral effect*. Zwróćmy też uwagę, jak duża jest różnica pomiędzy liczbą udostępnień powyższych artykułów.

Przyjrzyjmy się jeszcze co się stanie po odrzuceniu 100, 500 i 1000-cu największych obserwacji:



Rozkład zaczyna już być zauważalny, widzimy, że najpopularniejsze są artykuły od 1500 do 3000 udostępnień. Przy takiej liczbie odrzuconych obserwacji wartość maksymalna przewidywanej cechy to

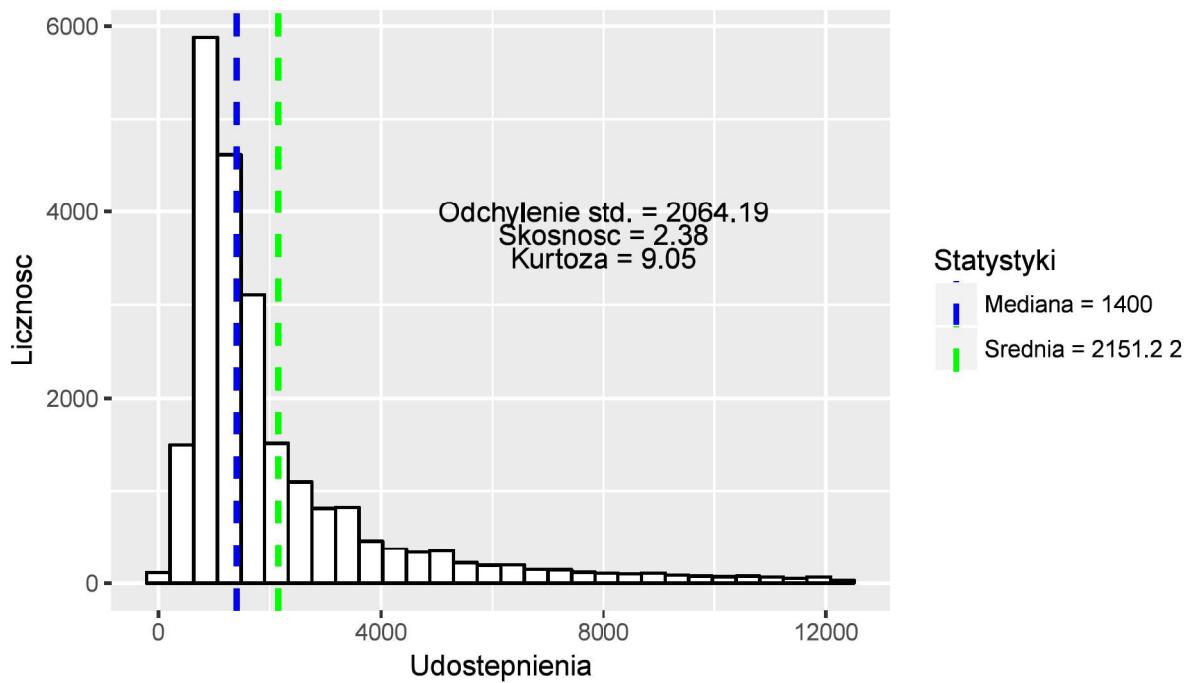
Dla porządku odrzućmy 500 największych rekordów:



Zbiór obcięty o 500 obserwacji może być już zbyt mocno okrojony, aby zbudować rozsądny model, ale pozwala nam on przyjrzeć się jak dokładnie wygląda rozkład liczby udostępnień dla mniejszych wartości. Weźmy pod uwagę nasz cel, na powyższym histogramie widzimy, że górną granicą udostępnień stosunkowo rzadko przekracza 10000 (dokładnie w 5.6% przypadków) gdyby udało nam się przesunąć ją np. do 15000 to byłby to duży sukces.

Dla porządku rozpatrzmy jeszcze najbardziej okrojony zbiór:

Histogram bez 1000 największych obserwacji

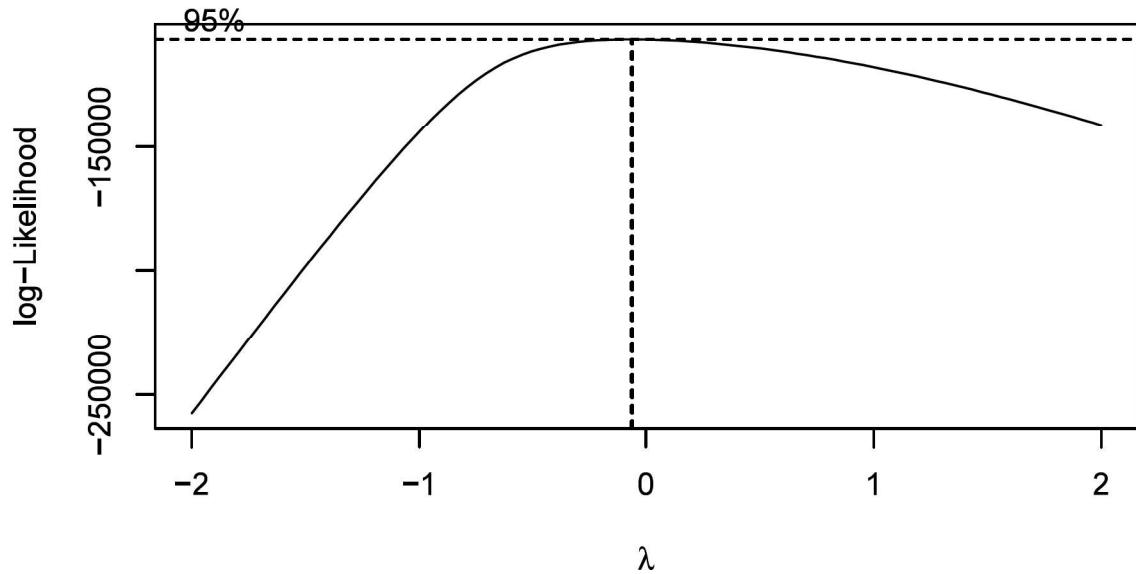


Powyżej widzimy histogram po odrzuceniu 4.2% obserwacji.

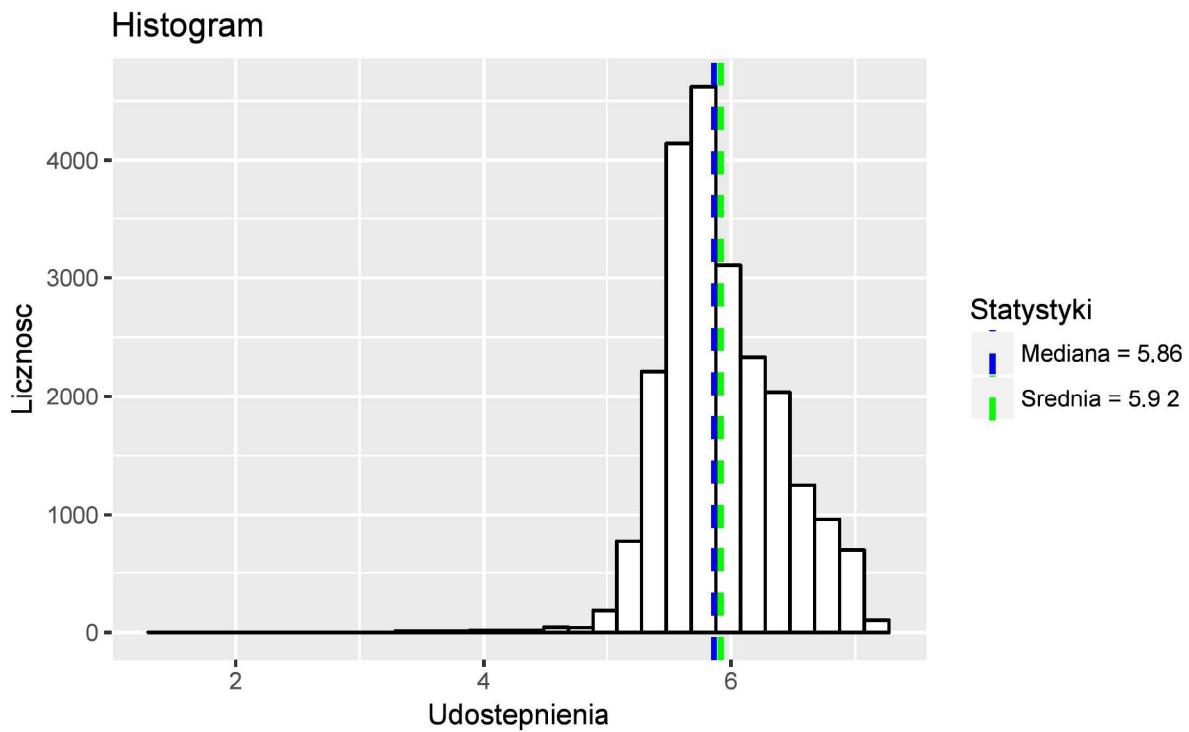
Ustalamy, że udrzucamy 100 największych obserwacji, jest to na tyle mała liczba, że można się im przyjrzeć ręcznie. Przeanalizować czy miały w sobie coś wyjątkowego, czy to zasługa tematyki lub “chwytnego pióra”.

Na końcu użyjemy transformacji Boxa-Coxa, najpierw sprawdźmy jaka wartość lamby najlepiej zsymetryzuje dane:

```
bc=MASS::boxcox(shares~1,data=popularityTrain)
```



Następnie narysujemy rozkład share'ów po transformacji:

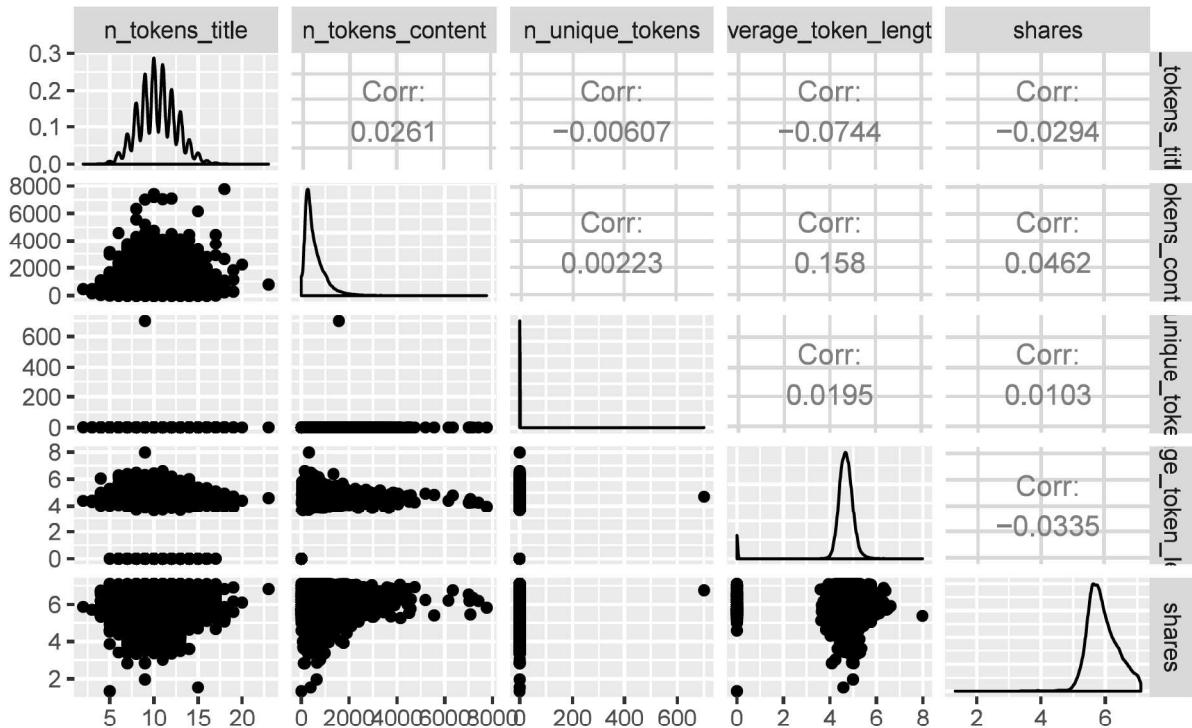


Wyrazy

Pierwsze kolumny zawierają informację o budowie artykułu, przypomnijmy:

```
...
5. n_tokens_title: Number of words in the title
6. n_tokens_content: Number of words in the content
7. n_unique_tokens: Rate of unique words in the content
8. n_non_stop_words: Rate of non-stop words in the content
9. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
...
14. average_token_length: Average length of the words in the content
...
```

Są to sześć faktów, bez analizy nacechowania samych słów, te informacje są ujęte w kolejnych kolumnach. Sprawdźmy jak wygląda ich rozkład:



Bardzo zastanawia obserwacja z dużą wartością **n_unique_tokens**, sprawdźmy ją:

Tablica 4: Dziesięć najpopularniejszych artykułów

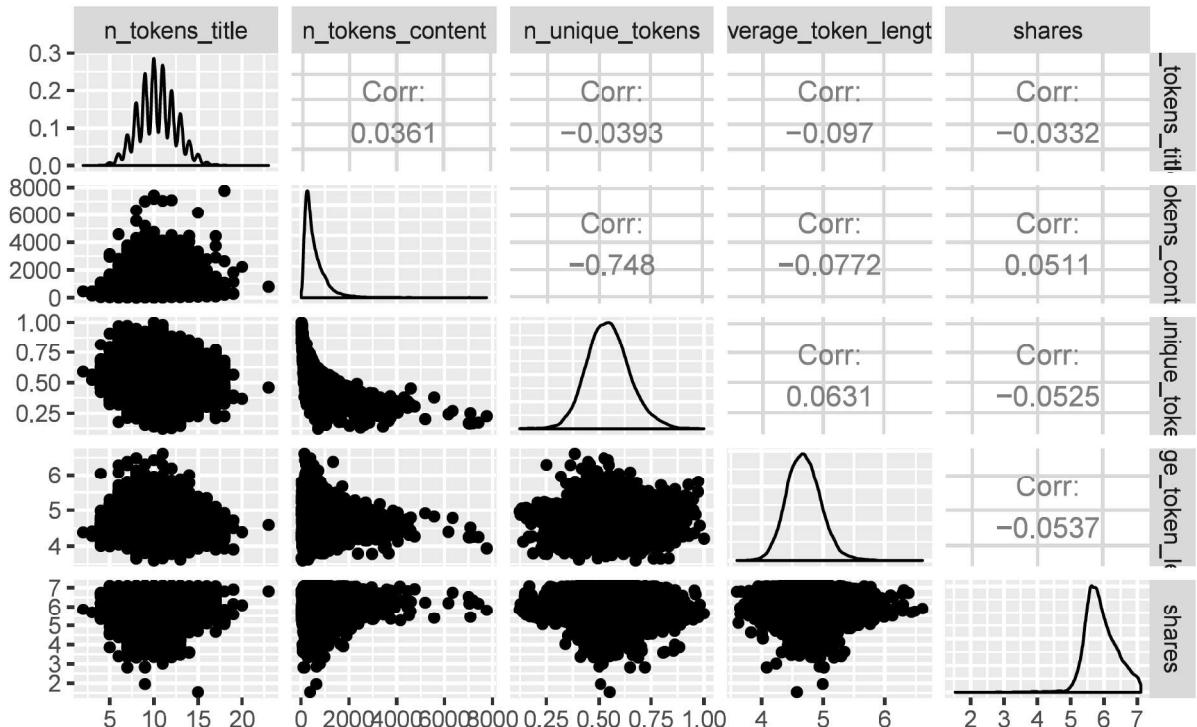
n_tokens_title	n_tokens_content	n_unique_tokens	average_token_length	shares
31038	9	1570	701.0000000	4.696178 6.751283
35377	10	24	1.0000000	4.208333 5.644102
3547	11	49	0.9795918	4.571429 6.121435
7515	7	37	0.9729729	5.810811 5.431371
27271	6	35	0.9714285	5.514286 5.636822
2223	10	25	0.9600000	4.880000 6.664811

Jak widać rekord numer **31038** jest niepoprawny, w takim razie usuwamy tę obserwację.

Pamiętajmy, że nasze dane zostały spreparowane przy pomocy metod *machine learningu*, więc mogły się tam wkrąć błędy, na poziomie samej analizy, ale też później, na poziomie przygotowania.

Usuniemy także rekordy, które posiadają odstające od średniej wartości, w naszym przypadku **average_token_length** większe od siedmiu.

Przyglądnijmy się rozkładom klasy po usunięciu, naszym zdaniem, błędnych lub odstających rekordów:



Golym okiem nie widać już większych anomalii. Spoglądając na najniższy wiersz powyżej kraty widać rozkład **share'ów** od elementów klasy, ciężko dopatrzeć się tutaj większych zależności... Widzimy jedynie, że rozkład jest skumulowany wokół średniej wartości każdego ze współczynników.

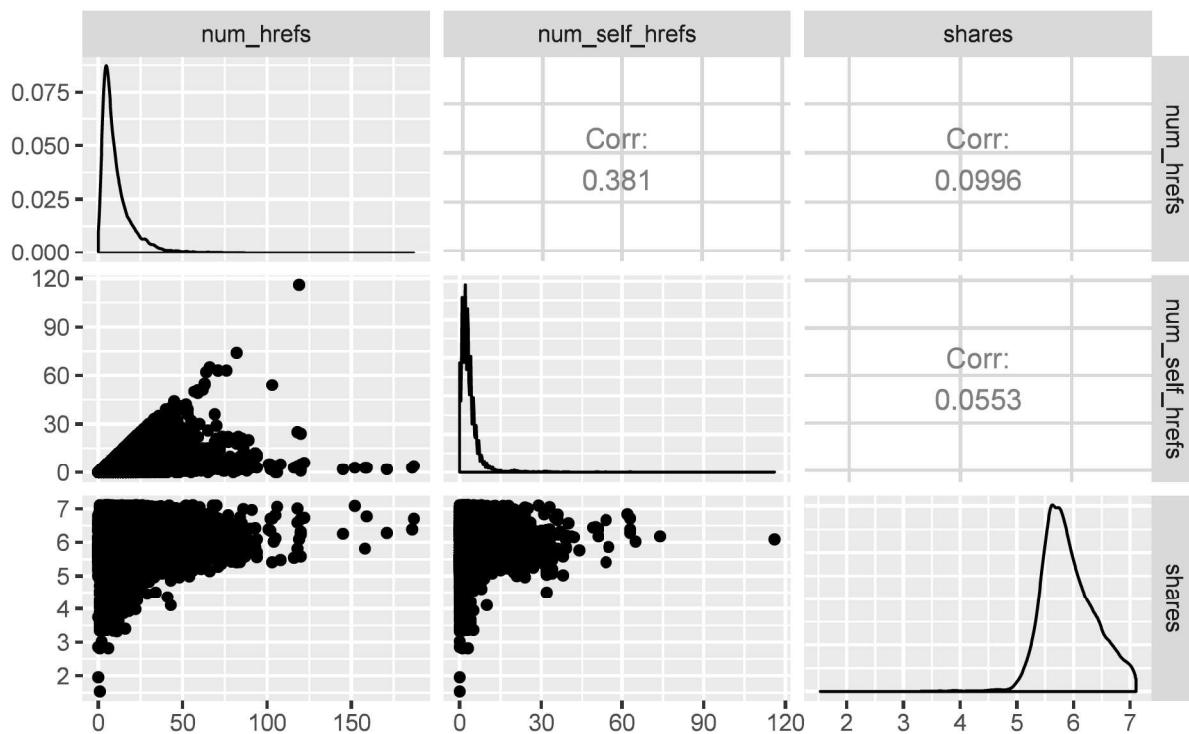
Odbośniki

Pięć kolumn niesie informacje o linkach w artykule:

```
...
10. num_hrefs: Number of links
11. num_self_hrefs: Number of links to other articles published by Mashable
...
26. self_reference_min_shares: Min. shares of referenced articles in Mashable
27. self_reference_max_shares: Max. shares of referenced articles in Mashable
28. self_reference_avg_shares: Avg. shares of referenced articles in Mashable
...
```

Linki

Sprawdźmy jak ma się rozkład dwóch pierwszych:



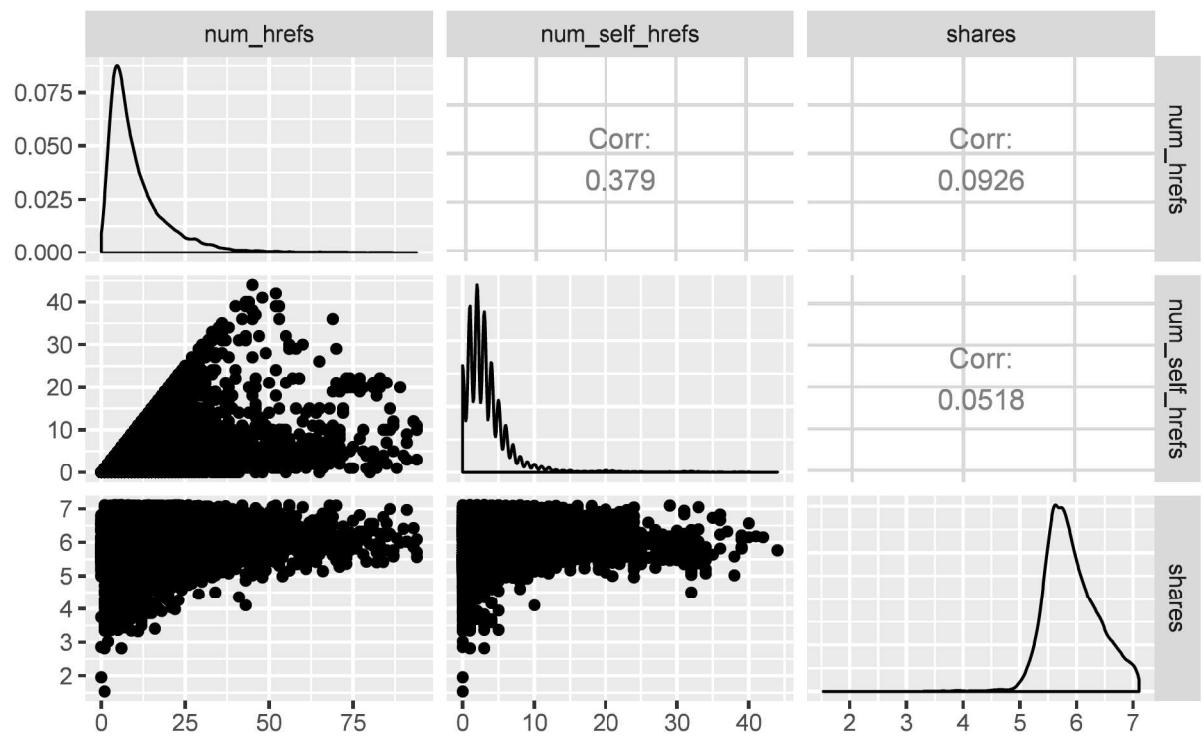
Oczywiście linki do Mashable powinny być podzbiorem linków w ogólności, sprawdźmy czy tak rzeczywiście jest:

```
subset(popularityTrain, num_href < num_self_href) [,c(4,52)]
```

```
## [1] url      shares
## <0 rows> (or 0-length row.names)
```

Wszystko się zgadza, tabela jest pusta.

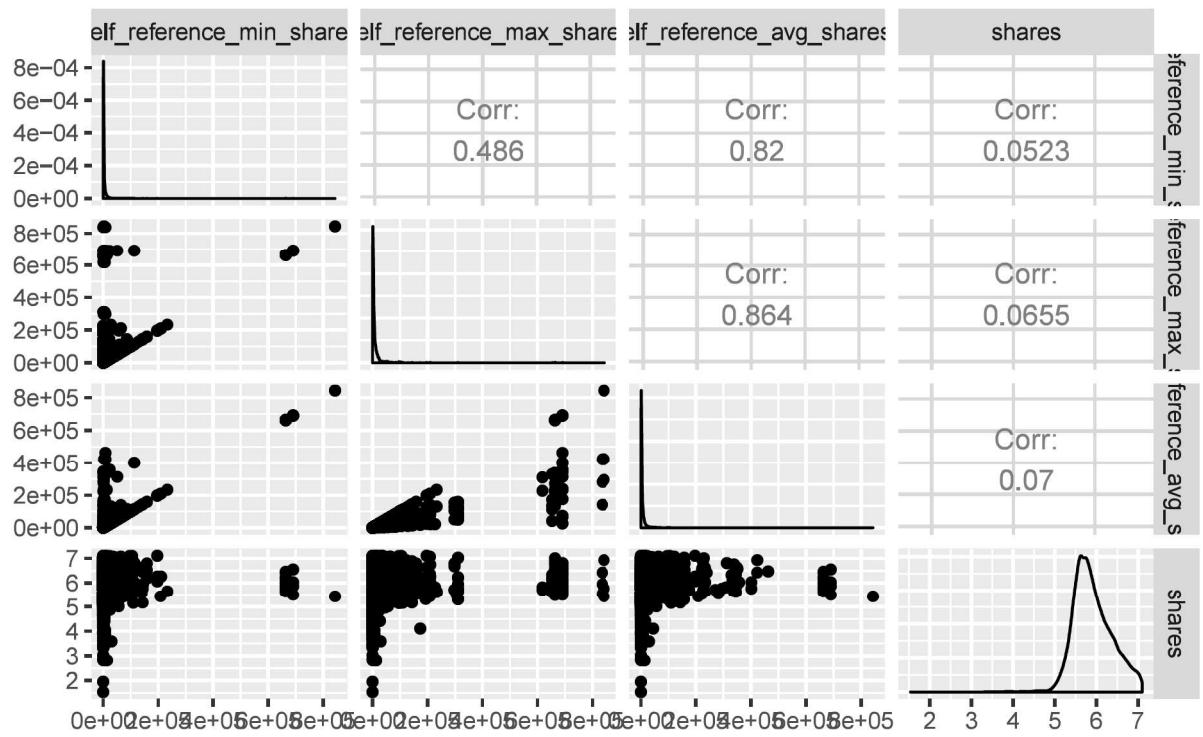
Z racji sporych obserwacji odstających usuniemy te największe, który umieszcza powyżej 100 linków w przeciętnym artykule? Usuniemy też te, które mają ponad 45 linków do Mashable.



Ponownie nie widzimy większych zależności pomiędzy liczbą odnośników, a liczbą share'ów...

Referencje

Teraz sprawdźmy, jak rozkładają się udostępnienia artykułów do których ten ukryty pod danym rekordem:



Oczywiście nie dziwi fakt, że obserwacje są bardzo mocno skorelowane, ale niestety jedynie pomiędzy sobą, korelacja z ilością udostępnien nie jest już tak widoczna. Jest jasne, jeśli ktoś natrafi na bardziej interesujący artykuł, to właśnie na nim się skupi. Gdybyśmy posiadali statystykę odwrotną tzn. z jakich artykułów możemy przejść od obserwowanego, wtedy, być może, moglibyśmy coś na tej podstawie wyciągnąć.

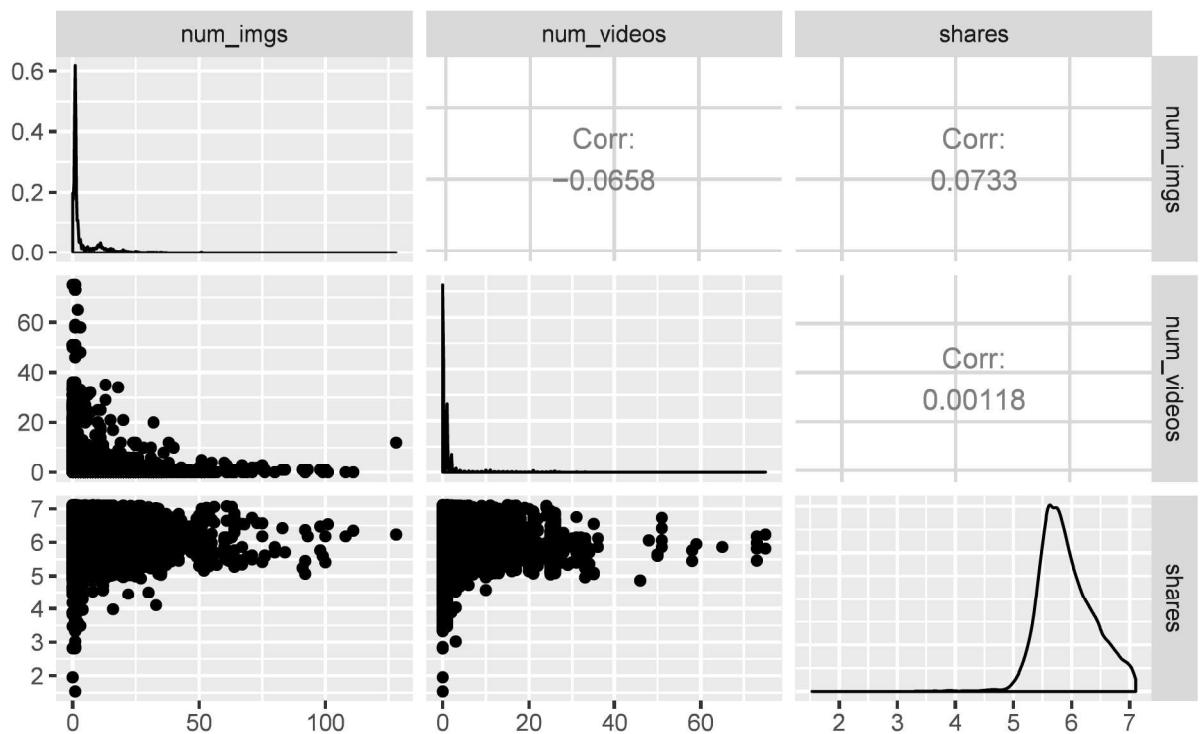
Media

Informacja o mediach użytych w artykule jest zapisana w dwóch kolumnach:

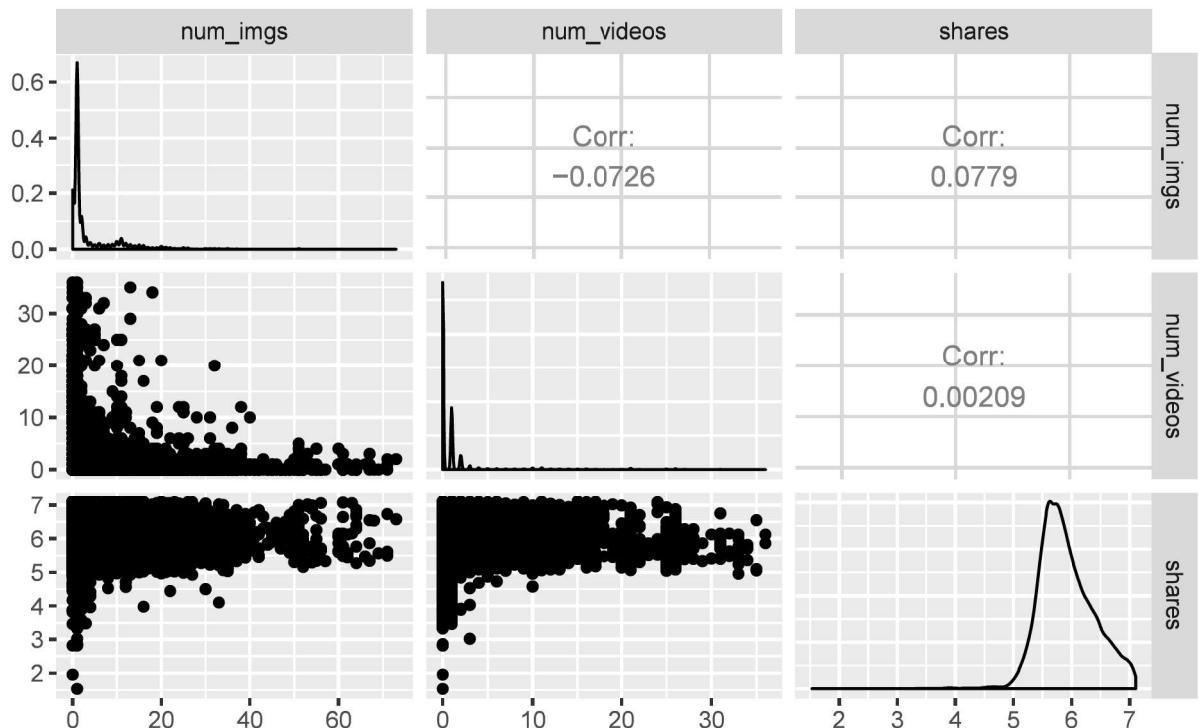
```
...
12. num_imgs: Number of images
13. num_videos: Number of videos
...

```

Sprawdźmy jak ma się ich rozkład:



Po raz kolejny ograniczymy nasz zbiór, więcej niż 40 wideo oraz więcej niż 75 zdjęć w artykule, to już bardzo dużo, dlatego “przytniemy” nasz zbiór powyżej tych wartości.



Jako, że dane pozostawiają wiele do życzenia, duża ilość obserwacji dla pewnych wielkości num_videos oraz num_imgs być jest konsekwencją błędów, ale nie możemy tego sprawdzić.

Nie widzimy dużej korelacji pomiędzy liczbą medii, a liczbą udostępnień.

Czas

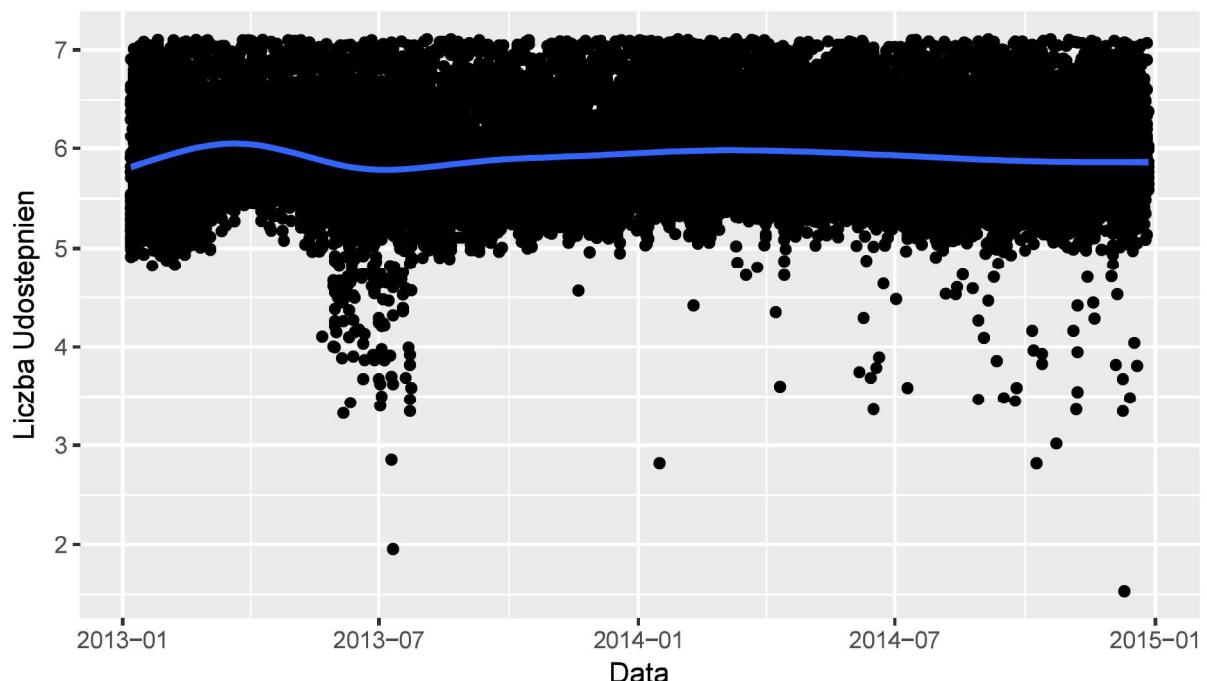
W kolejnej klasie przeanalizujemy cztery kolumny z dwoma informacjami:

1. year: Year of publication
2. month: Month of publication
3. day : Day of publication
- ...
29. weekday: Weekday of publishing
30. is_weekend: Was the article published on the weekend?
- ...

Przyjrzyjmy się jak wygląda zależności udostępnień od tych informacji:

Data

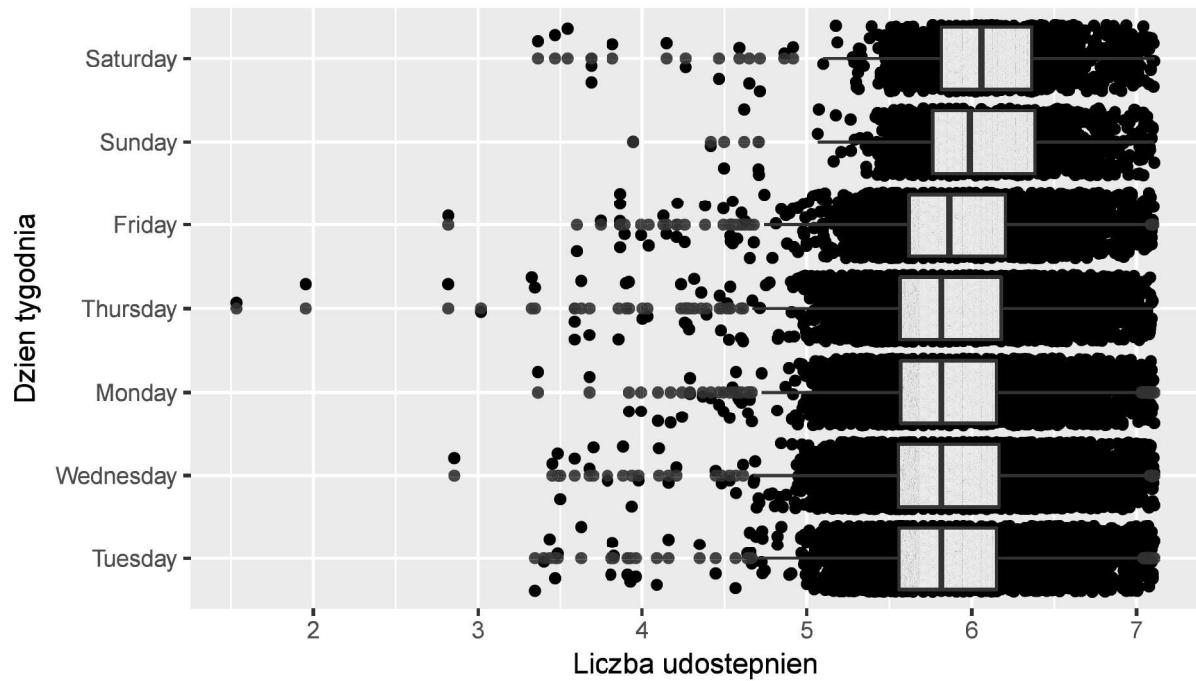
Data vs Udostepniania



Nie widzimy żadne zależności, mogłoby się wydawać, że z czasem serwis Mashable.com stawał się coraz popularniejszy, jednakże nie jest to prawda. Dane równomiernie się rozkładają.

Dzień tygodnia

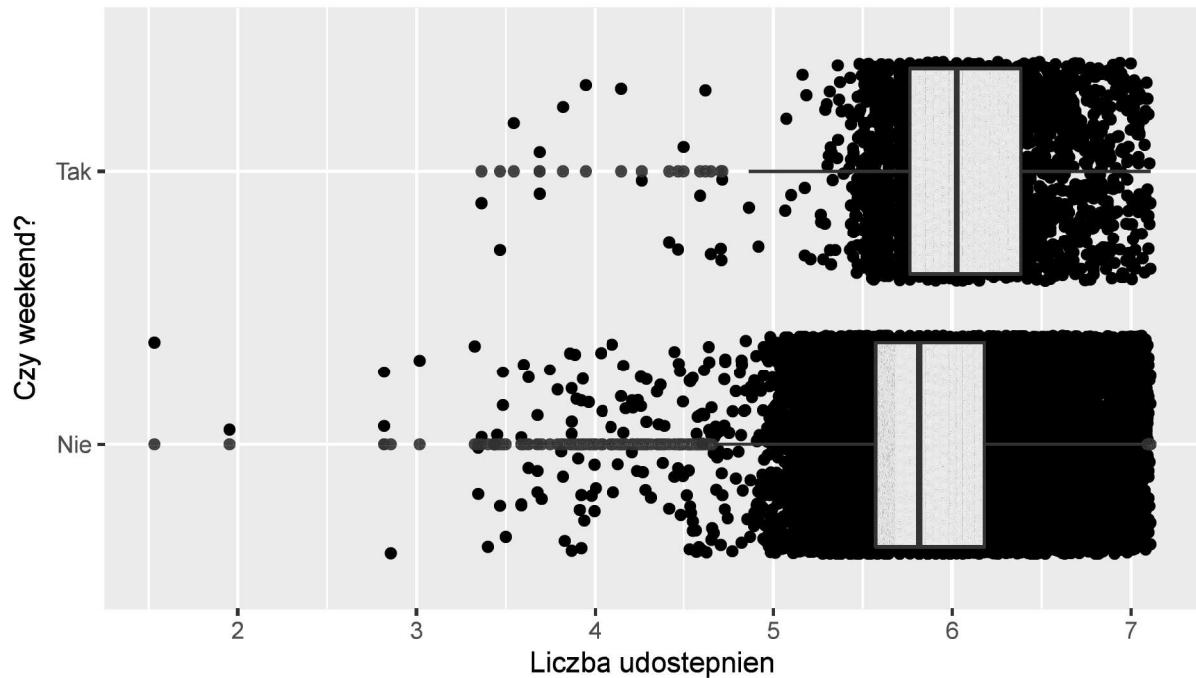
Dzien tygodnia vs Udostepnienia



Powyższy wykres jest uporządkowany, tzn. najwyżej są dni o największej, średniej ilości share'ów. Zgodne jest to z intuicją, że w weekendy ludzie spędzają więcej przed komputerem i ilość udostępnień rośnie. Najniższa jest w środku tygodnia, lecz to wtedy pojawia się więcej obserwacji odstających.

Sprawdźmy jeszcze jak się kształtują dane przy podziale na weekend oraz reszte tygodnia:

Weekend vs Udostepnienia



Tutaj potwierdzają się nasze wczesniejsze obserwacje, możemy je podsumować:

- Jeśli chcemy uzyskać wysoką liczbę udostępnień artykułu, ale nie jest on sensacją, to opublikujmy go w weekend
- Jeśli natomiast jest to sensacja albo coś, co może stać się “wirusem” sieci społecznościowych, to opublikujmy to w tygodniu

Póki co nie możemy nic więcej stwierdzić.

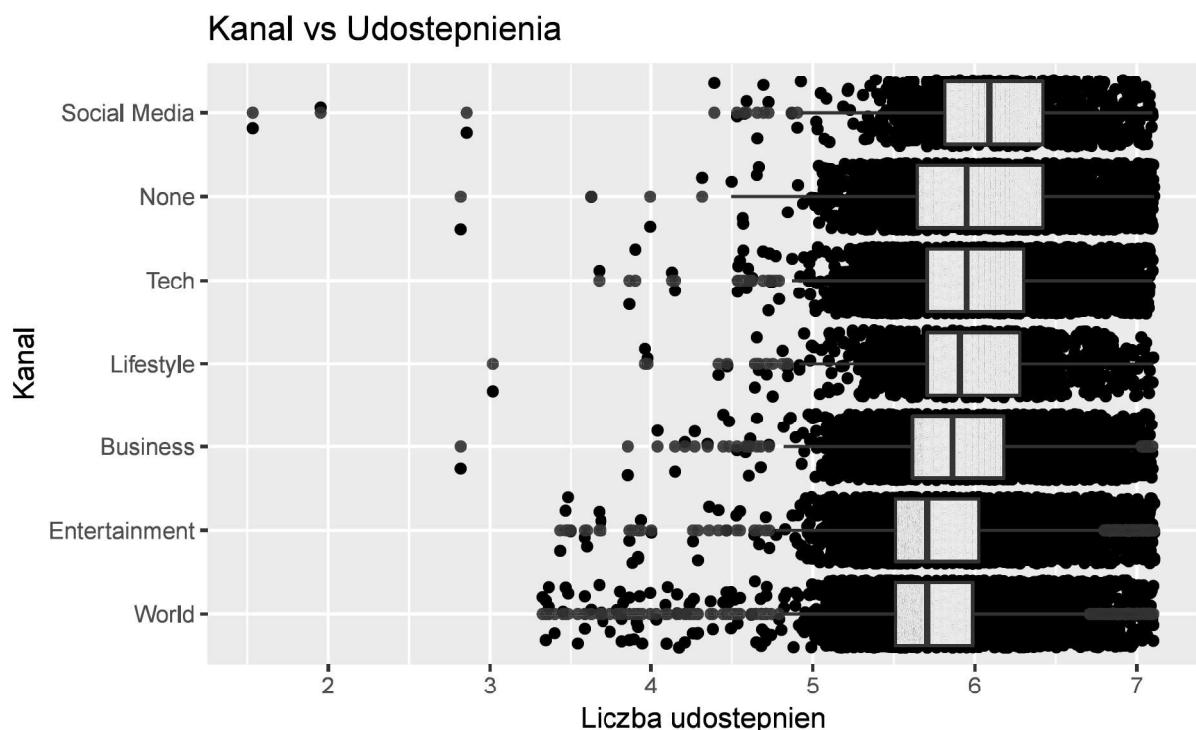
Słowa kluczowe

Jedna z dwóch, najbardziej licznych klas:

...
15. num_keywords: Number of keywords in the metadata
16. data_channel: 'Lifestyle', 'Entertainment', 'Business', 'Social Media', 'Tech', 'World' or NA?
17. kw_min_min: Worst keyword (min. shares)
18. kw_max_min: Worst keyword (max. shares)
19. kw_avg_min: Worst keyword (avg. shares)
20. kw_min_max: Best keyword (min. shares)
21. kw_max_max: Best keyword (max. shares)
22. kw_avg_max: Best keyword (avg. shares)
23. kw_min_avg: Avg. keyword (min. shares)
24. kw_max_avg: Avg. keyword (max. shares)
25. kw_avg_avg: Avg. keyword (avg. shares)
...

Kanał

Zaczniemy od analizy data_channel:

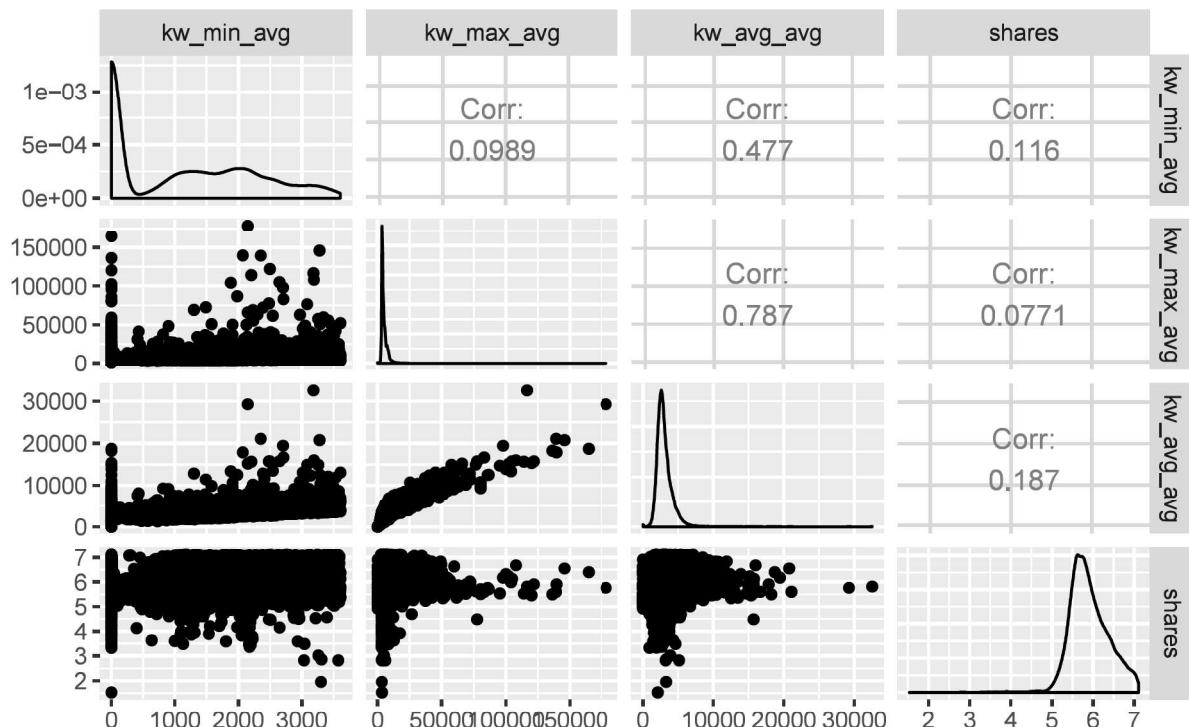


Najpopularniejszym kanałem stał się... brak kanału! Takie mamy dane, tyle one nam mówią. Prawdopodobnie, gdy autor artykułu stwierdzał, że będzie on popularny to wrzucał go na stronę główną, bez kanału. Możemy potwierdzić, że jeżeli tak faktycznie było, to redaktorzy mają poprawne intuicje. Wiele więcej tutaj nie widzimy.

Słowa kluczowe

Przyjrzymy się rozkładowi pozostałych zmiennych, lecz najpierw zrobimy to dla "średniego" słowa kluczowego i wszystkich poziomów udostępnień, następnie dla wszystkich poziomów słów kluczowych, ale "średnich" udostępnień.

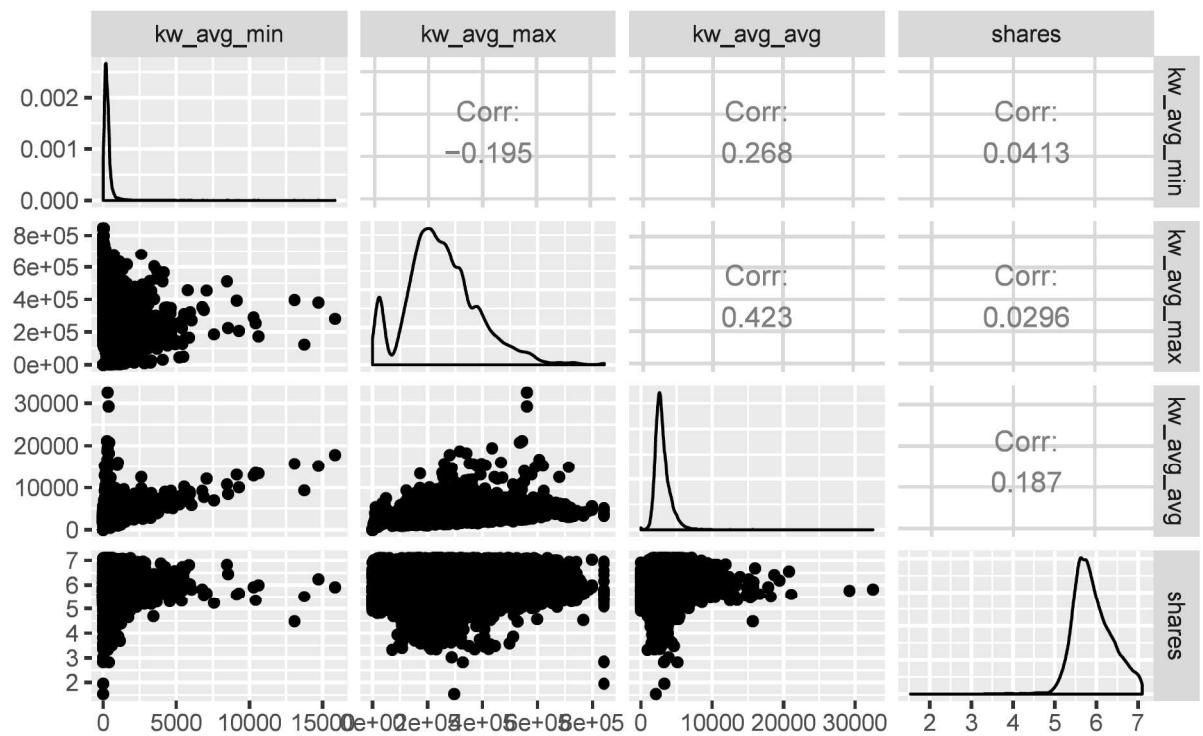
```
...
23. kw_min_avg: Avg. keyword (min. shares)
24. kw_max_avg: Avg. keyword (max. shares)
25. kw_avg_avg: Avg. keyword (avg. shares)
...
```



Zgodnie z tradycją: korelacja pomiędzy trzema pierwszymi zmiennymi wysoka, ale pomiędzy nimi, a ilością share'ów znikoma...

Spójrzmy teraz na to z innej strony:

```
...
19. kw_avg_min: Worst keyword (avg. shares)
22. kw_avg_max: Best keyword (avg. shares)
25. kw_avg_avg: Avg. keyword (avg. shares)
...
```

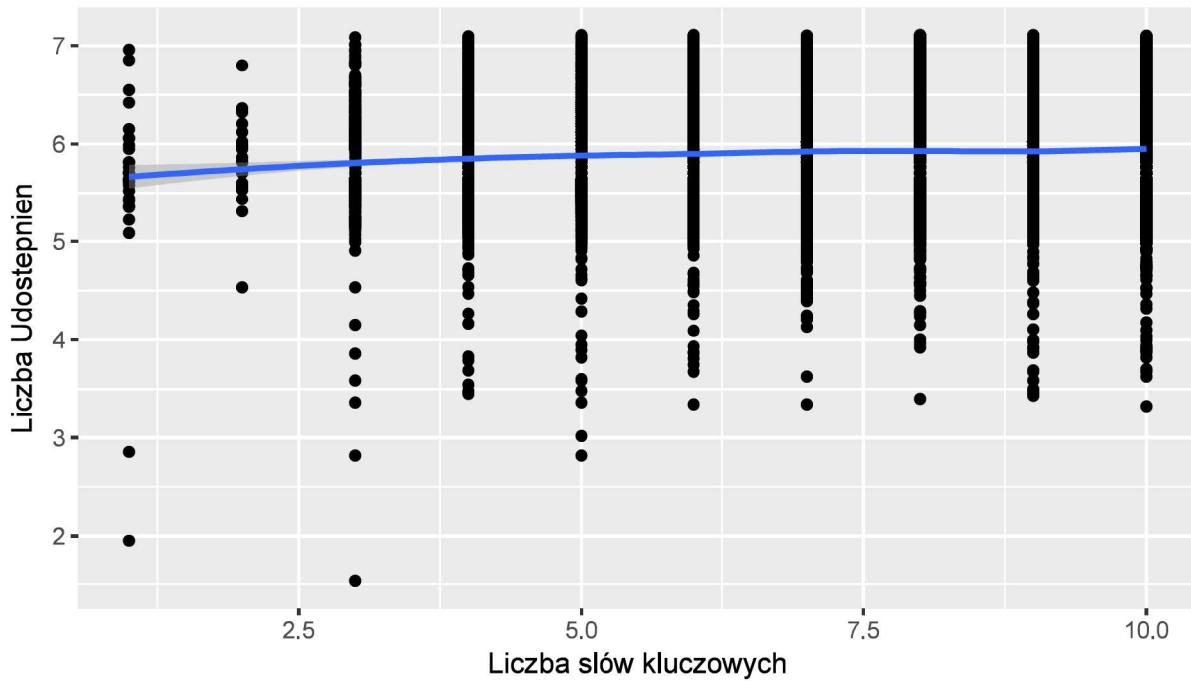


Ponownie, niewiele widzimy, możemy zaobserwować, jak wygląda rozkład, ale jakakolwiek zależność pomiędzy zmiennymi wyjaśniającymi, a wyjaśnianą jest nie widoczna gołym okiem.

Pozostała nam jeszcze do sprawdzenia jak na popularność wpływa liczba słów kluczowych:

```
...
15. num_keywords: Number of keywords in the metadata
...
...
```

Slowa kluczowe vs Udostepniania



Widać, że co do zasady, artykuły z małą ilością słów kluczowych nie są ekstremalnie popularne, ale ta zależność ginie w ogromie obserwacji jaką posiadamy. Jest to zależność liniowa o bardzo niskim współczynniku kierunkowym.

NLP

Pozostała nam do analizy największa klasa dot. języka:

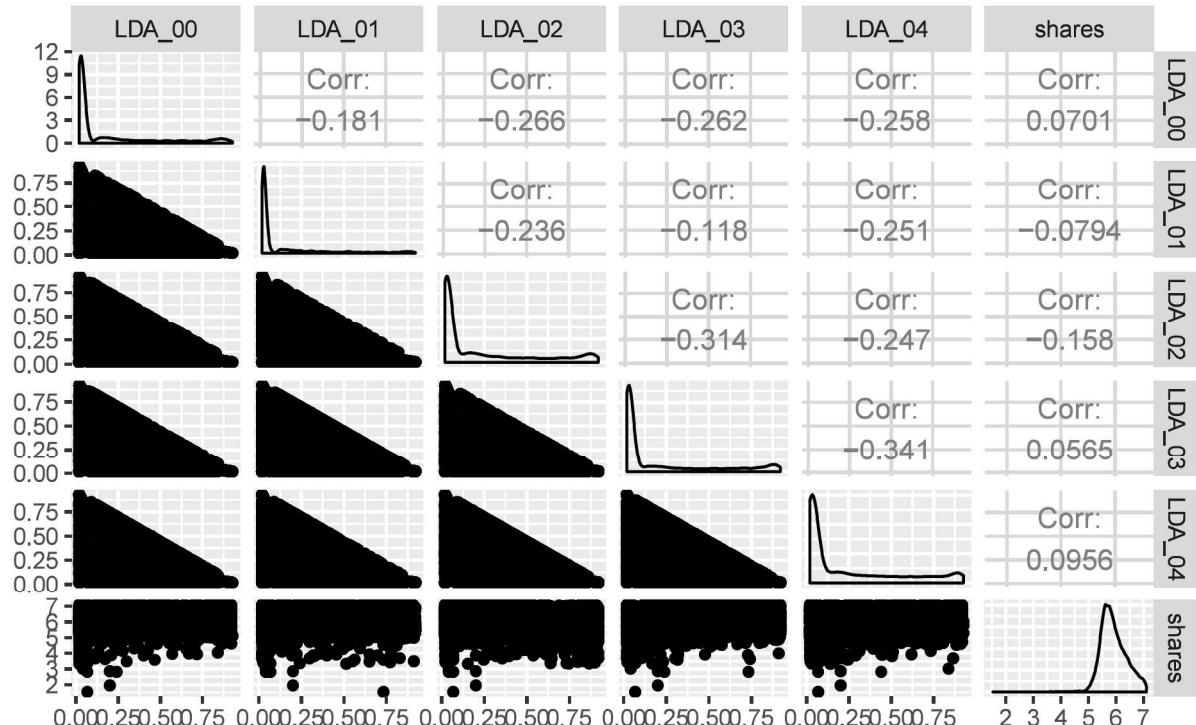
- ```
...
31. LDA_00: Closeness to LDA topic 0
32. LDA_01: Closeness to LDA topic 1
33. LDA_02: Closeness to LDA topic 2
34. LDA_03: Closeness to LDA topic 3
35. LDA_04: Closeness to LDA topic 4
36. global_subjectivity: Text subjectivity
37. global_sentiment_polarity: Text sentiment polarity
38. global_rate_positive_words: Rate of positive words in the content
39. global_rate_negative_words: Rate of negative words in the content
40. rate_positive_words: Rate of positive words among non-neutral tokens
41. rate_negative_words: Rate of negative words among non-neutral tokens
42. avg_positive_polarity: Avg. polarity of positive words
43. min_positive_polarity: Min. polarity of positive words
44. max_positive_polarity: Max. polarity of positive words
45. avg_negative_polarity: Avg. polarity of negative words
46. min_negative_polarity: Min. polarity of negative words
47. max_negative_polarity: Max. polarity of negative words
48. title_subjectivity: Title subjectivity
49. title_sentiment_polarity: Title polarity
50. abs_title_subjectivity: Absolute subjectivity level
```

51. abs\_title\_sentiment\_polarity: Absolute polarity level

...

## LDA

LDA - Latent Dirichlet allocation, to model, który jest wykorzystywany w przetwarzaniu masywnym języka. Autor zbioru danych nie umieścił dokładnej informacji, co kryje się pod tą informacją. Klasycznie przyjdźmy do wykresy *ggpairs*:

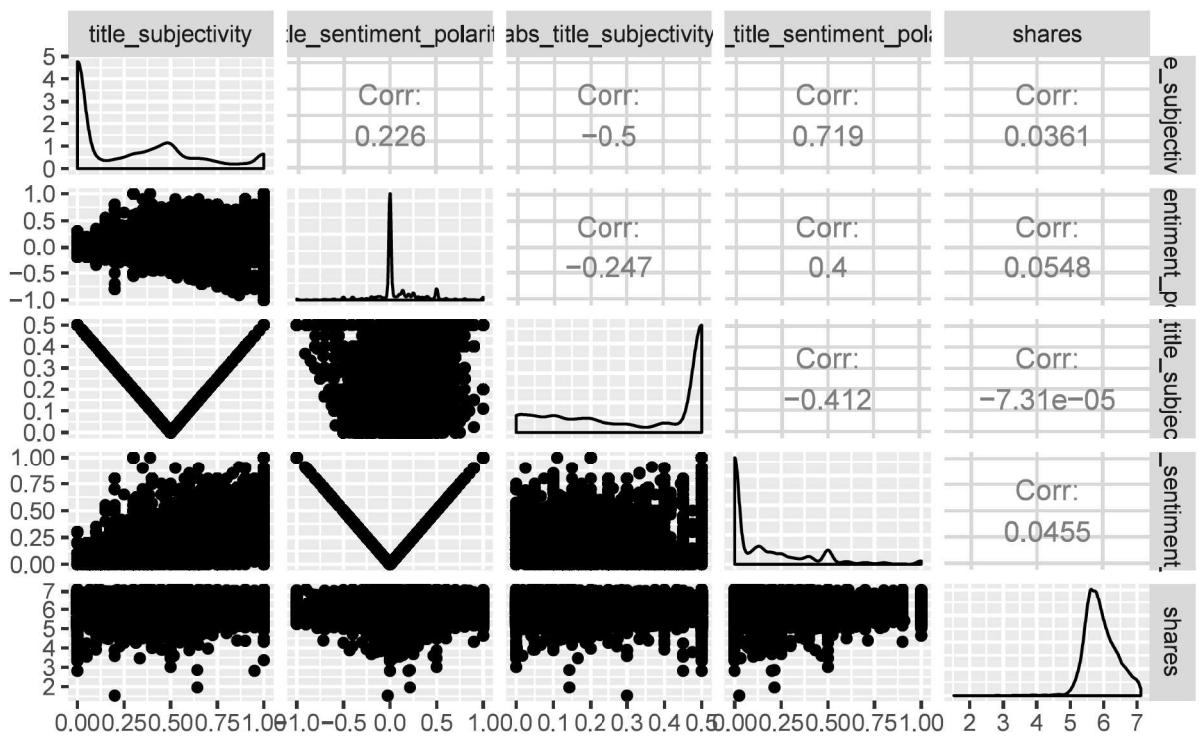


Klasycznie, nie jesteśmy w stanie gołym okiem znaleźć żadnych zależności... Możemy jednak dzięki nim zobaczyć jak rozkładają się informacje w danej kolumnie.

## Wydzwięk

Najpierw przyjrzymy się kolumnom 36-41:





Widzimy jak rozkładają się poszczególne zmienne i jakie przyjmują wartości, lecz zależności nie są takie oczywiste.

## Podsumowanie

Analizując wysoką korelację pewnych zmiennych objaśniających, a także niesioną przez nie informacje tworzymy nową ramkę danych okrojoną z niepotrzebnych (naszym zdaniem) kolumn:

Number of Attributes: 28 (27 predictive attributes, 1 goal field)

Attribute Information:

1. month: Month of publication
2. n\_tokens\_title: Number of words in the title
3. n\_tokens\_content: Number of words in the content
4. n\_unique\_tokens: Rate of unique words in the content
5. n\_non\_stop\_words: Rate of non-stop words in the content
6. n\_non\_stop\_unique\_tokens: Rate of unique non-stop words in the content
7. num\_hrefs: Number of links
8. num\_imgs: Number of images
9. average\_token\_length: Average length of the words in the content
10. num\_keywords: Number of keywords in the metadata
11. data\_channel: 'Lifestyle', 'Entertainment', 'Business', 'Social Media', 'Tech', 'World' or NA?
12. kw\_avg\_max: Best keyword (avg. shares)
13. kw\_avg\_avg: Avg. keyword (avg. shares)
14. self\_reference\_avg\_shares: Avg. shares of referenced articles in Mashable
15. is\_weekend: Was the article published on the weekend?
16. LDA\_00: Closeness to LDA topic 0
17. LDA\_01: Closeness to LDA topic 1
18. LDA\_02: Closeness to LDA topic 2

19. LDA\_03: Closeness to LDA topic 3
20. LDA\_04: Closeness to LDA topic 4
21. global\_subjectivity: Text subjectivity
22. global\_rate\_positive\_words: Rate of positive words in the content
23. global\_rate\_negative\_words: Rate of negative words in the content
24. avg\_positive\_polarity: Avg. polarity of positive words
25. avg\_negative\_polarity: Avg. polarity of negative words
26. abs\_title\_subjectivity: Absolute subjectivity level
27. abs\_title\_sentiment\_polarity: Absolute polarity level
28. shares: Number of shares (target)

Jednakże cały czas zachowujemy starą ramkę (pomniejszoną o adres url, dzień i rok), może się okazać, że nasze przeczucia były błędne.

## Model

Zacznijmy budowanie modeli:

```
popularityLog=lm(shares~., data=popularityTrain)
summary(popularityLog)

##
Call:
lm(formula = shares ~ ., data = popularityTrain)
##
Residuals:
Min 1Q Median 3Q Max
-4.2768 -0.2804 -0.0600 0.2496 1.4565
##
Coefficients: (3 not defined because of singularities)
Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.742e+00 2.644e-01 21.716 < 2e-16 ***
month02 -1.891e-02 1.497e-02 -1.263 0.206510
month03 4.197e-02 1.825e-02 2.300 0.021450 *
month04 5.724e-02 1.587e-02 3.607 0.000311 ***
month05 -1.059e-03 1.616e-02 -0.066 0.947748
month06 -8.945e-02 1.588e-02 -5.632 1.80e-08 ***
month07 -8.022e-02 1.623e-02 -4.944 7.71e-07 ***
month08 -5.552e-02 1.646e-02 -3.372 0.000747 ***
month09 -6.629e-02 1.647e-02 -4.026 5.69e-05 ***
month10 -6.188e-02 1.604e-02 -3.858 0.000115 ***
month11 -3.131e-02 1.661e-02 -1.884 0.059533 .
month12 -5.479e-02 1.689e-02 -3.244 0.001180 **
n_tokens_title 2.838e-03 1.465e-03 1.938 0.052687 .
n_tokens_content 2.523e-05 1.175e-05 2.147 0.031836 *
n_unique_tokens -5.391e-03 9.840e-02 -0.055 0.956313
n_non_stop_words NA NA NA
n_non_stop_unique_tokens -1.347e-01 8.363e-02 -1.611 0.107162
num_hrefs 1.483e-03 3.858e-04 3.845 0.000121 ***
num_self_hrefs -2.706e-03 9.611e-04 -2.815 0.004878 **
num_imgs 4.109e-04 5.108e-04 0.804 0.421224
num_videos 1.424e-04 9.430e-04 0.151 0.879976
average_token_length -5.001e-02 1.245e-02 -4.015 5.95e-05 ***
num_keywords 7.850e-03 1.908e-03 4.114 3.91e-05 ***
data_channelEntertainment -1.658e-02 1.856e-02 -0.893 0.371677
data_channelLifestyle 1.097e-02 1.888e-02 0.581 0.561197
data_channelNone 6.915e-02 1.982e-02 3.488 0.000487 ***
data_channelSocial Media 1.971e-01 1.606e-02 12.271 < 2e-16 ***
data_channelTech 1.579e-01 1.684e-02 9.378 < 2e-16 ***
data_channelWorld 1.607e-02 1.817e-02 0.884 0.376487
kw_min_min 2.872e-04 1.083e-04 2.651 0.008024 **
kw_max_min 6.042e-06 3.772e-06 1.602 0.109236
kw_avg_min -4.657e-05 2.469e-05 -1.887 0.059213 .
kw_min_max -2.754e-07 6.194e-08 -4.446 8.80e-06 ***
kw_max_max -1.310e-08 3.715e-08 -0.353 0.724439
kw_avg_max -9.971e-08 4.381e-08 -2.276 0.022855 *
kw_min_avg -1.320e-05 3.962e-06 -3.332 0.000862 ***
kw_max_avg -2.156e-05 1.301e-06 -16.574 < 2e-16 ***
```

```

kw_avg_avg 1.635e-04 7.573e-06 21.583 < 2e-16 ***
self_reference_min_shares -5.353e-07 4.198e-07 -1.275 0.202252
self_reference_max_shares -2.738e-07 2.323e-07 -1.179 0.238473
self_reference_avg_sharess 1.489e-06 5.974e-07 2.493 0.012659 *
weekdayMonday -3.089e-02 1.062e-02 -2.910 0.003622 **
weekdaySaturday 1.337e-01 1.435e-02 9.315 < 2e-16 ***
weekdaySunday 1.315e-01 1.387e-02 9.484 < 2e-16 ***
weekdayThursday -4.871e-02 1.036e-02 -4.701 2.60e-06 ***
weekdayTuesday -5.881e-02 1.033e-02 -5.691 1.28e-08 ***
weekdayWednesday -4.681e-02 1.028e-02 -4.556 5.24e-06 ***
is_weekend1 NA NA NA NA
LDA_00 1.235e-01 2.323e-02 5.316 1.07e-07 ***
LDA_01 -1.144e-01 2.586e-02 -4.424 9.75e-06 ***
LDA_02 -8.793e-02 2.330e-02 -3.773 0.000162 ***
LDA_03 -1.053e-01 2.465e-02 -4.272 1.94e-05 ***
LDA_04 NA NA NA NA
global_subjectivity 1.054e-01 4.309e-02 2.447 0.014428 *
global_sentiment_polarity -5.551e-02 8.602e-02 -0.645 0.518712
global_rate_positive_words 2.152e-01 3.665e-01 0.587 0.557185
global_rate_negative_words -1.209e+00 7.120e-01 -1.698 0.089525 .
rate_positive_words 5.542e-02 2.537e-01 0.218 0.827072
rate_negative_words 3.461e-02 2.565e-01 0.135 0.892671
avg_positive_polarity 5.200e-02 7.017e-02 0.741 0.458663
min_positive_polarity -2.030e-01 5.885e-02 -3.450 0.000562 ***
max_positive_polarity -3.922e-02 2.178e-02 -1.800 0.071807 .
avg_negative_polarity 1.524e-02 6.389e-02 0.238 0.811524
min_negative_polarity -2.406e-02 2.324e-02 -1.035 0.300711
max_negative_polarity 9.134e-03 5.293e-02 0.173 0.862992
title_subjectivity 3.548e-02 1.434e-02 2.474 0.013370 *
title_sentiment_polarity 3.102e-02 1.295e-02 2.395 0.016638 *
abs_title_subjectivity 6.403e-02 1.885e-02 3.397 0.000682 ***
abs_title_sentiment_polarity 5.208e-03 2.055e-02 0.253 0.799949

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 0.4368 on 21818 degrees of freedom
Multiple R-squared: 0.1435, Adjusted R-squared: 0.1409
F-statistic: 56.22 on 65 and 21818 DF, p-value: < 2.2e-16
popularityLogBIS=lm(shares~.,data=popularityTrainBIS)
summary(popularityLogBIS)

##
Call:
lm(formula = shares ~ ., data = popularityTrainBIS)
##
Residuals:
Min 1Q Median 3Q Max
-4.3496 -0.2830 -0.0610 0.2553 1.4326
##
Coefficients: (2 not defined because of singularities)
Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.858e+00 7.837e-02 74.750 < 2e-16 ***
month02 -2.006e-03 1.505e-02 -0.133 0.893940
month03 6.770e-02 1.537e-02 4.404 1.07e-05 ***


```

```

month04 5.222e-02 1.511e-02 3.457 0.000548 ***
month05 -2.368e-03 1.534e-02 -0.154 0.877361
month06 -9.160e-02 1.520e-02 -6.026 1.71e-09 ***
month07 -8.827e-02 1.520e-02 -5.809 6.39e-09 ***
month08 -6.125e-02 1.538e-02 -3.982 6.85e-05 ***
month09 -7.068e-02 1.545e-02 -4.574 4.82e-06 ***
month10 -6.713e-02 1.496e-02 -4.489 7.21e-06 ***
month11 -3.615e-02 1.566e-02 -2.308 0.021001 *
month12 -5.832e-02 1.598e-02 -3.648 0.000264 ***
n_tokens_title 2.058e-03 1.476e-03 1.394 0.163241
n_tokens_content 1.921e-05 1.129e-05 1.701 0.088941 .
n_unique_tokens -8.775e-02 9.231e-02 -0.951 0.341806
n_non_stop_words NA NA NA NA
n_non_stop_unique_tokens -7.316e-02 8.105e-02 -0.903 0.366753
num_hrefs 1.654e-03 3.663e-04 4.514 6.38e-06 ***
num_imgs 6.501e-04 4.966e-04 1.309 0.190524
average_token_length -4.580e-02 1.246e-02 -3.677 0.000236 ***
num_keywords 5.918e-03 1.793e-03 3.301 0.000964 ***
data_channelEntertainment -1.026e-02 1.843e-02 -0.557 0.577774
data_channelLifestyle 5.004e-02 1.881e-02 2.661 0.007800 **
data_channelNone 1.211e-01 1.974e-02 6.137 8.53e-10 ***
data_channelSocial Media 2.199e-01 1.583e-02 13.892 < 2e-16 ***
data_channelTech 1.637e-01 1.687e-02 9.703 < 2e-16 ***
data_channelWorld 1.705e-02 1.813e-02 0.940 0.347208
kw_avg_max -1.224e-08 3.202e-08 -0.382 0.702385
kw_avg_avg 5.894e-05 3.283e-06 17.952 < 2e-16 ***
self_reference_avg_shares 5.918e-07 1.260e-07 4.695 2.68e-06 ***
is_weekend1 1.755e-01 9.041e-03 19.407 < 2e-16 ***
LDA_00 1.274e-01 2.334e-02 5.458 4.88e-08 ***
LDA_01 -1.017e-01 2.598e-02 -3.913 9.13e-05 ***
LDA_02 -1.102e-01 2.338e-02 -4.715 2.44e-06 ***
LDA_03 -6.663e-02 2.445e-02 -2.725 0.006429 **
LDA_04 NA NA NA
global_subjectivity 1.530e-01 4.183e-02 3.658 0.000255 ***
global_rate_positive_words 2.935e-01 2.093e-01 1.402 0.160830
global_rate_negative_words -9.274e-01 3.087e-01 -3.004 0.002664 **
avg_positive_polarity -8.957e-02 3.934e-02 -2.277 0.022791 *
avg_negative_polarity -1.949e-02 2.767e-02 -0.704 0.481179
abs_title_subjectivity 5.014e-02 1.787e-02 2.805 0.005030 **
abs_title_sentiment_polarity 5.570e-02 1.499e-02 3.715 0.000203 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 0.441 on 21843 degrees of freedom
Multiple R-squared: 0.1258, Adjusted R-squared: 0.1242
F-statistic: 78.55 on 40 and 21843 DF, p-value: < 2.2e-16

```

Widzimy, że mimo naszych intuicji większy model daje lepsze dopasowanie (jeżeli za miarę potraktujemy  $R^2$ )...

Odtąd będziemy posługiwać się nazwami \*popularityLog oraz popularityLogBIS\*\*.

## ANOVA

Przeprowadźmy ANOVA'ę dla sprawdzenia modeli:

```
Analysis of Variance Table
##
Model 1: shares ~ 1
Model 2: shares ~ month + n_tokens_title + n_tokens_content + n_unique_tokens +
n_non_stop_words + n_non_stop_unique_tokens + num_hrefs +
num_self_hrefs + num_imgs + num_videos + average_token_length +
num_keywords + data_channel + kw_min_min + kw_max_min + kw_avg_min +
kw_min_max + kw_max_max + kw_avg_max + kw_min_avg + kw_max_avg +
kw_avg_avg + self_reference_min_shares + self_reference_max_shares +
self_reference_avg_shares + weekday + is_weekend + LDA_00 +
LDA_01 + LDA_02 + LDA_03 + LDA_04 + global_subjectivity +
global_sentiment_polarity + global_rate_positive_words +
global_rate_negative_words + rate_positive_words + rate_negative_words +
avg_positive_polarity + min_positive_polarity + max_positive_polarity +
avg_negative_polarity + min_negative_polarity + max_negative_polarity +
title_subjectivity + title_sentiment_polarity + abs_title_subjectivity +
abs_title_sentiment_polarity
Res.Df RSS Df Sum of Sq F Pr(>F)
1 21883 4859.7
2 21818 4162.5 65 697.13 56.216 < 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Analysis of Variance Table
##
Model 1: shares ~ 1
Model 2: shares ~ month + n_tokens_title + n_tokens_content + n_unique_tokens +
n_non_stop_words + n_non_stop_unique_tokens + num_hrefs +
num_imgs + average_token_length + num_keywords + data_channel +
kw_avg_max + kw_avg_avg + self_reference_avg_shares + is_weekend +
LDA_00 + LDA_01 + LDA_02 + LDA_03 + LDA_04 + global_subjectivity +
global_rate_positive_words + global_rate_negative_words +
avg_positive_polarity + avg_negative_polarity + abs_title_subjectivity +
abs_title_sentiment_polarity
Res.Df RSS Df Sum of Sq F Pr(>F)
1 21883 4859.7
2 21843 4248.5 40 611.16 78.555 < 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

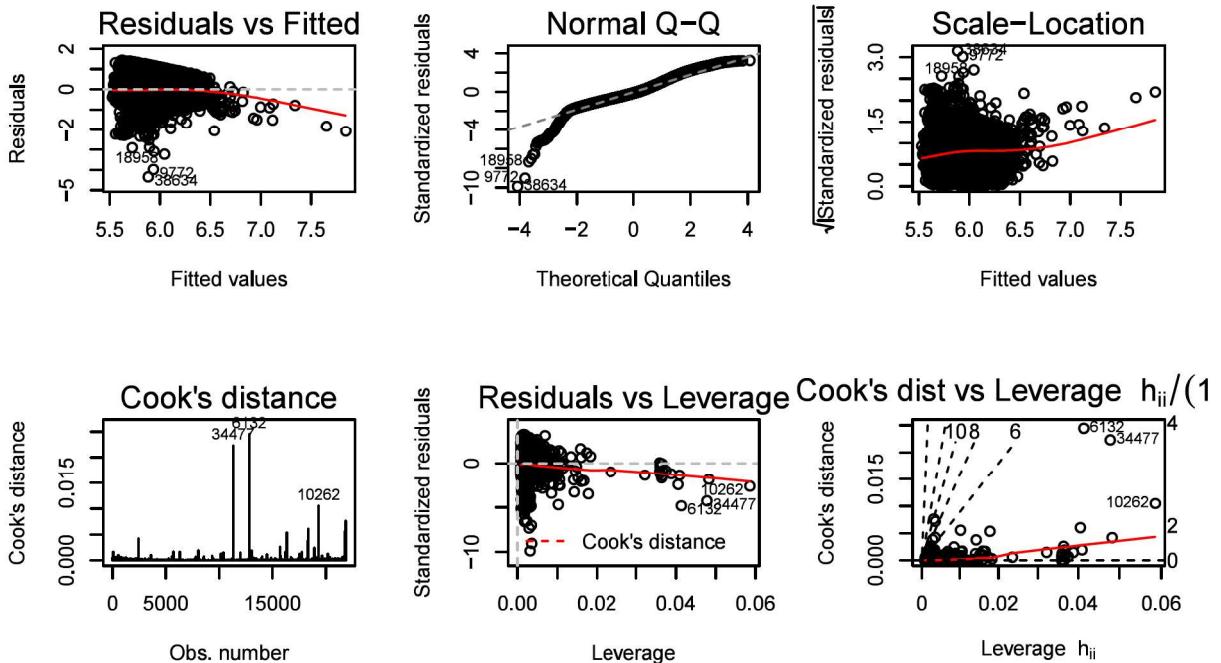
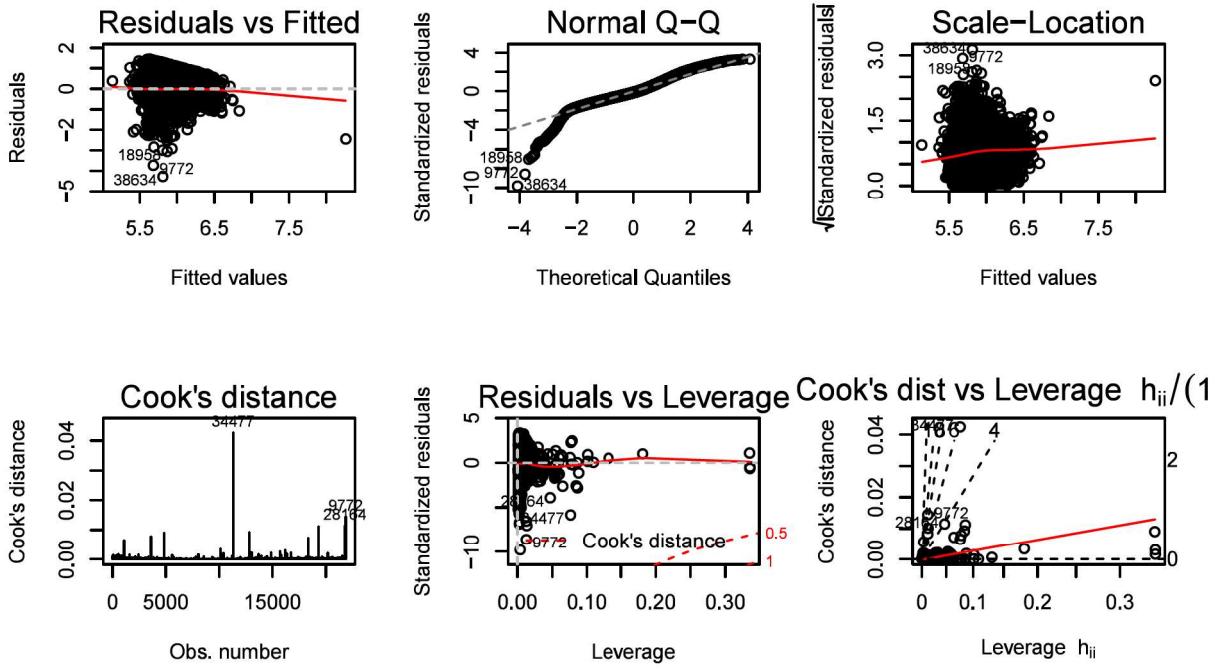
Jak widzimy RRS dla skonstruowanych modelu jest mniejsze niż dla modeli zerowego, wartości statystyki duża, P-wartości małe, więc nasze modele zdają się być lepsze, od modeli zerowych.

## Wykresy diagnostyczne

Przejdzmy do analizy wykresów diagnostycznych.

```
par(mfrow=c(2,3))

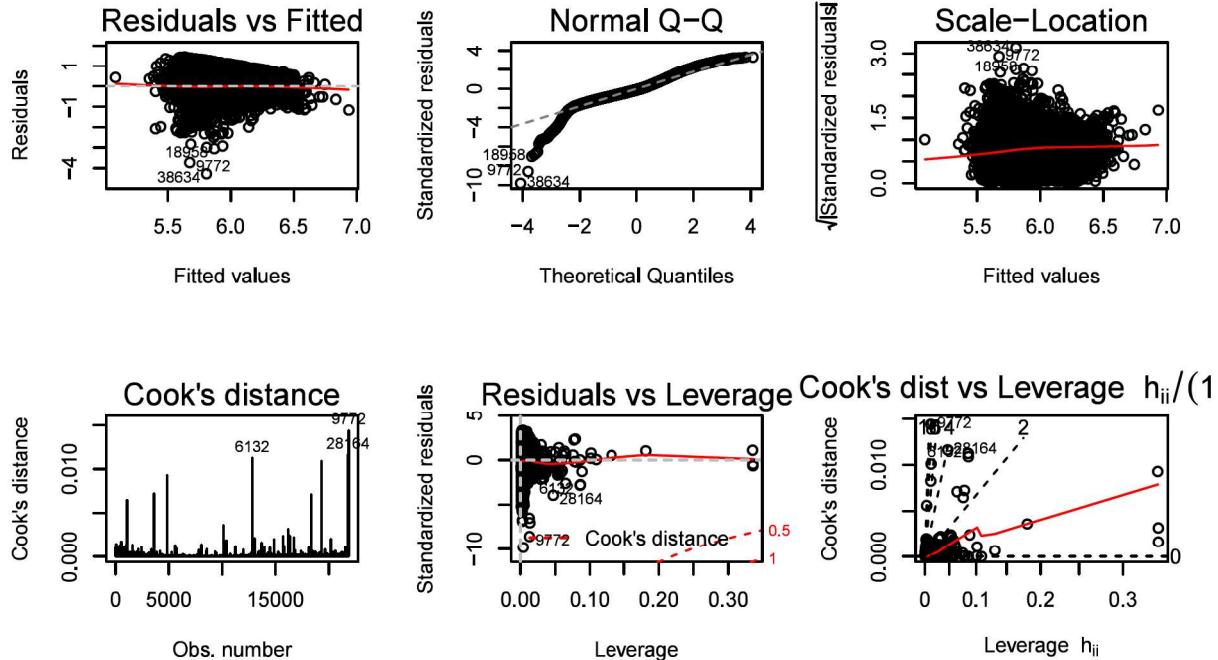
plot(popularityLog, which = 1:6)
```



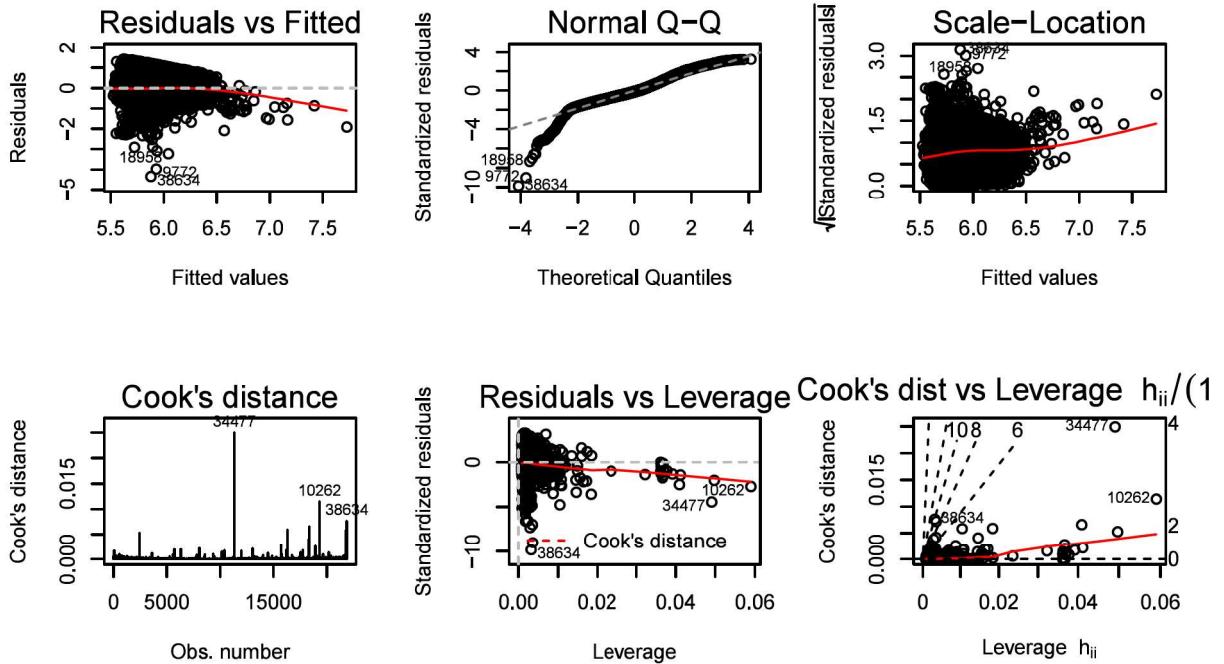
Zauważmy, że kilka obserwacji odstaje, usuniemy je z naszych danych i znów narysujemy wykresy diagnostyczne:

```
par(mfrow=c(2,3))

plot(popularityLog, which = 1:6)
```



```
plot(popularityLogBIS, which = 1:6)
```



Jak widać, nie jest idealnie, nie jest nawet dobrze, szczególnie jeżeli chodzi o rozkład residiuów, ale niewiele więcej możemy na tę chwilę zrobić.

## Testy

Model **popularityLog**:

```
bptest(popularityLog)

##
studentized Breusch-Pagan test
##
data: popularityLog
BP = 402.4, df = 65, p-value < 2.2e-16
gqttest(popularityLog)

##
Goldfeld-Quandt test
##
data: popularityLog
GQ = 0.60241, df1 = 10873, df2 = 10872, p-value = 1
alternative hypothesis: variance increases from segment 1 to 2
dwtest(popularityLog, order.by=fitted(popularityLog))

##
Durbin-Watson test
##
data: popularityLog
DW = 1.9806, p-value = 0.7603
```

```

alternative hypothesis: true autocorrelation is greater than 0

Sprawdzanie normalności dla tak duże próby niekoniecznie ma sens, dlatego ją pominiemy. Z testów na homoskedastyczność jestnie test BP odrzuca hipoteze zerową.

Model popularityLogBIS

bptest(popularityLogBIS)

##
studentized Breusch-Pagan test
##
data: popularityLogBIS
BP = 326.31, df = 40, p-value < 2.2e-16

gqtest(popularityLogBIS)

##
Goldfeld-Quandt test
##
data: popularityLogBIS
GQ = 0.60763, df1 = 10897, df2 = 10897, p-value = 1
alternative hypothesis: variance increases from segment 1 to 2

dwttest(popularityLogBIS, order.by=fitted(popularityLogBIS))

##
Durbin-Watson test
##
data: popularityLogBIS
DW = 1.9677, p-value = 0.7603
alternative hypothesis: true autocorrelation is greater than 0

```

Wartości statystyk i testów są praktycznie takie same jak dla modelu **popularityLog**, wnioski też są podobne

## BIC

Jednym ze sposobów na poszukiwanie najlepszego modelu w regresji liniowej jest Bayesian Information Criterion, zwane w skrócie BIC. Nie będziemy tutaj zagłębiać się w podstawy teoretyczne tej metody, zastosujemy ją i zobaczymy jakie przynosi efekty, poniżej zobaczymy podsumowanie obydwu modeli po działaniu algorytmu zachłanego na podstawie BIC:

```

summary(LogBIC)

##
Call:
lm(formula = shares ~ month + n_non_stop_unique_tokens + num_hrefs +
average_token_length + num_keywords + data_channel + kw_min_min +
kw_min_max + kw_min_avg + kw_max_avg + kw_avg_avg + self_reference_avg_shares +
weekday + LDA_00 + LDA_01 + LDA_02 + LDA_03 + min_positive_polarity +
global_subjectivity, data = popularityTrain)
##
Residuals:
Min 1Q Median 3Q Max
-4.2744 -0.2810 -0.0605 0.2519 1.4521
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
```

```

(Intercept) 5.845e+00 6.722e-02 86.954 < 2e-16 ***
month02 -1.865e-02 1.493e-02 -1.250 0.211430
month03 3.195e-02 1.522e-02 2.099 0.035787 *
month04 5.015e-02 1.561e-02 3.213 0.001315 **
month05 -8.084e-03 1.587e-02 -0.510 0.610380
month06 -9.462e-02 1.567e-02 -6.038 1.58e-09 ***
month07 -8.569e-02 1.557e-02 -5.505 3.73e-08 ***
month08 -6.052e-02 1.576e-02 -3.841 0.000123 ***
month09 -6.999e-02 1.579e-02 -4.433 9.32e-06 ***
month10 -6.491e-02 1.531e-02 -4.241 2.24e-05 ***
month11 -3.532e-02 1.590e-02 -2.222 0.026304 *
month12 -5.767e-02 1.625e-02 -3.548 0.000388 ***
n_non_stop_unique_tokens -2.056e-01 3.326e-02 -6.181 6.49e-10 ***
num_hrefs 1.489e-03 3.405e-04 4.371 1.24e-05 ***
average_token_length -5.346e-02 1.148e-02 -4.656 3.25e-06 ***
num_keywords 8.900e-03 1.769e-03 5.032 4.89e-07 ***
data_channelEntertainment -1.130e-02 1.805e-02 -0.626 0.531371
data_channelLifestyle 1.791e-02 1.840e-02 0.973 0.330320
data_channelNone 6.135e-02 1.960e-02 3.131 0.001746 **
data_channelSocial Media 1.986e-01 1.539e-02 12.909 < 2e-16 ***
data_channelTech 1.592e-01 1.662e-02 9.579 < 2e-16 ***
data_channelWorld 2.371e-02 1.775e-02 1.336 0.181705
kw_min_min 3.629e-04 5.776e-05 6.283 3.39e-10 ***
kw_min_max -3.117e-07 5.756e-08 -5.415 6.21e-08 ***
kw_min_avg -1.573e-05 3.945e-06 -3.987 6.70e-05 ***
kw_max_avg -2.145e-05 1.205e-06 -17.800 < 2e-16 ***
kw_avg_avg 1.649e-04 7.120e-06 23.167 < 2e-16 ***
self_reference_avg_sharess 6.842e-07 1.250e-07 5.472 4.49e-08 ***
weekdayMonday -3.226e-02 1.061e-02 -3.041 0.002360 **
weekdaySaturday 1.340e-01 1.434e-02 9.346 < 2e-16 ***
weekdaySunday 1.324e-01 1.382e-02 9.575 < 2e-16 ***
weekdayThursday 4.921e-02 1.036e-02 -4.752 2.03e-06 ***
weekdayTuesday 5.855e-02 1.033e-02 -5.670 1.45e-08 ***
weekdayWednesday 4.671e-02 1.027e-02 -4.547 5.46e-06 ***
LDA_00 1.200e-01 2.309e-02 5.200 2.01e-07 ***
LDA_01 -1.143e-01 2.574e-02 -4.441 9.02e-06 ***
LDA_02 -8.554e-02 2.312e-02 -3.700 0.000216 ***
LDA_03 -1.107e-01 2.418e-02 -4.576 4.77e-06 ***
min_positive_polarity -2.132e-01 4.577e-02 -4.658 3.21e-06 ***
global_subjectivity 1.226e-01 3.682e-02 3.329 0.000874 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 0.4369 on 21843 degrees of freedom
Multiple R-squared: 0.142, Adjusted R-squared: 0.1405
F-statistic: 92.72 on 39 and 21843 DF, p-value: < 2.2e-16
summary(LogBICBIS)

##
Call:
lm(formula = shares ~ month + n_unique_tokens + num_hrefs + average_token_length +
num_keywords + data_channel + kw_avg_avg + self_reference_avg_sharess +
is_weekend + LDA_00 + global_subjectivity + LDA_04, data = popularityTrainBIS)
##

```

```

Residuals:
Min 1Q Median 3Q Max
-4.3597 -0.2832 -0.0617 0.2567 1.4407
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.789e+00 6.331e-02 91.436 < 2e-16 ***
month02 -2.698e-03 1.504e-02 -0.179 0.85762
month03 6.301e-02 1.525e-02 4.132 3.62e-05 ***
month04 4.739e-02 1.484e-02 3.194 0.00141 **
month05 -7.793e-03 1.502e-02 -0.519 0.60393
month06 -9.835e-02 1.481e-02 -6.643 3.15e-11 ***
month07 -9.431e-02 1.471e-02 -6.413 1.46e-10 ***
month08 -6.633e-02 1.492e-02 -4.444 8.86e-06 ***
month09 -7.404e-02 1.496e-02 -4.950 7.46e-07 ***
month10 -7.120e-02 1.444e-02 -4.930 8.29e-07 ***
month11 -4.188e-02 1.508e-02 -2.777 0.00549 **
month12 -6.297e-02 1.544e-02 -4.079 4.53e-05 ***
n_unique_tokens -2.242e-01 3.325e-02 -6.744 1.58e-11 ***
num_hrefs 2.005e-03 3.418e-04 5.865 4.54e-09 ***
average_token_length -4.417e-02 1.170e-02 -3.776 0.00016 ***
num_keywords 6.560e-03 1.646e-03 3.987 6.72e-05 ***
data_channelEntertainment -2.257e-03 1.682e-02 -0.134 0.89326
data_channelLifestyle 4.800e-02 1.848e-02 2.597 0.00941 **
data_channelNone 1.352e-01 1.824e-02 7.413 1.28e-13 ***
data_channelSocial Media 2.238e-01 1.544e-02 14.497 < 2e-16 ***
data_channelTech 1.646e-01 1.673e-02 9.837 < 2e-16 ***
data_channelWorld 8.733e-03 1.629e-02 0.536 0.59194
kw_avg_avg 6.240e-05 3.025e-06 20.632 < 2e-16 ***
self_reference_avg_shares 6.382e-07 1.264e-07 5.049 4.49e-07 ***
is_weekend1 1.751e-01 9.023e-03 19.402 < 2e-16 ***
LDA_00 2.226e-01 2.177e-02 10.224 < 2e-16 ***
global_subjectivity 1.556e-01 3.688e-02 4.220 2.46e-05 ***
LDA_04 9.792e-02 1.997e-02 4.902 9.55e-07 ***
##

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 0.4411 on 21852 degrees of freedom
Multiple R-squared: 0.125, Adjusted R-squared: 0.1239
F-statistic: 115.6 on 27 and 21852 DF, p-value: < 2.2e-16

```

Ilość zmiennych nam się znacząco zmniejszyła, a  $R^2$  pozostało praktycznie bez zmian, jest to o tyle dobre, że mocno uprościliśmy modele.

AIC

Modyfikacja BIC jest AIC, które różni się jedynie funkcją, której argumentem maksymalizującego szukamy.

```
summary(LogAIC)
```

```

Call:
lm(formula = shares ~ month + n_tokens_title + n_tokens_content +
n non stop unique tokens + num hrefs + num self hrefs + average token length +
```

```

num_keywords + data_channel + kw_min_min + kw_max_min + kw_avg_min +
kw_min_max + kw_avg_max + kw_min_avg + kw_max_avg + kw_avg_avg +
self_reference_avg_sharess + weekday + LDA_00 + LDA_01 +
LDA_02 + LDA_03 + global_subjectivity + global_rate_negative_words +
min_positive_polarity + max_positive_polarity + title_subjectivity +
title_sentiment_polarity + abs_title_subjectivity, data = popularityTrain)
##
Residuals:
Min 1Q Median 3Q Max
-4.2847 -0.2807 -0.0606 0.2501 1.4495
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.786e+00 7.923e-02 73.028 < 2e-16 ***
month02 -2.047e-02 1.492e-02 -1.372 0.170147
month03 3.663e-02 1.540e-02 2.379 0.017364 *
month04 5.508e-02 1.572e-02 3.505 0.000458 ***
month05 -3.226e-03 1.600e-02 -0.202 0.840234
month06 -9.148e-02 1.572e-02 -5.821 5.94e-09 ***
month07 -8.279e-02 1.564e-02 -5.293 1.22e-07 ***
month08 -5.777e-02 1.584e-02 -3.648 0.000265 ***
month09 -6.785e-02 1.588e-02 -4.272 1.94e-05 ***
month10 -6.265e-02 1.541e-02 -4.065 4.81e-05 ***
month11 -3.335e-02 1.603e-02 -2.080 0.037548 *
month12 -5.727e-02 1.635e-02 -3.503 0.000461 ***
n_tokens_title 2.917e-03 1.461e-03 1.997 0.045891 *
n_tokens_content 3.169e-05 9.543e-06 3.320 0.000901 ***
n_non_stop_unique_tokens -1.430e-01 3.864e-02 -3.701 0.000215 ***
num_hrefs 1.464e-03 3.790e-04 3.864 0.000112 ***
num_self_hrefs -2.619e-03 9.304e-04 -2.815 0.004886 **
average_token_length -5.011e-02 1.190e-02 -4.210 2.56e-05 ***
num_keywords 7.171e-03 1.894e-03 3.786 0.000154 ***
data_channelEntertainment -1.957e-02 1.842e-02 -1.062 0.288118
data_channelLifestyle 5.629e-03 1.887e-02 0.298 0.765407
data_channelNone 6.236e-02 1.972e-02 3.162 0.001567 **
data_channelSocial Media 1.928e-01 1.599e-02 12.052 < 2e-16 ***
data_channelTech 1.550e-01 1.682e-02 9.218 < 2e-16 ***
data_channelWorld 1.303e-02 1.806e-02 0.722 0.470415
kw_min_min 3.272e-04 6.690e-05 4.891 1.01e-06 ***
kw_max_min 6.153e-06 3.765e-06 1.634 0.102202
kw_avg_min -5.165e-05 2.465e-05 -2.095 0.036141 *
kw_min_max -2.666e-07 6.150e-08 -4.335 1.46e-05 ***
kw_avg_max -1.241e-07 4.180e-08 -2.970 0.002980 **
kw_min_avg -1.626e-05 3.979e-06 -4.086 4.40e-05 ***
kw_max_avg -2.222e-05 1.296e-06 -17.148 < 2e-16 ***
kw_avg_avg 1.730e-04 7.667e-06 22.565 < 2e-16 ***
self_reference_avg_sharess 7.097e-07 1.253e-07 5.662 1.52e-08 ***
weekdayMonday -3.044e-02 1.060e-02 -2.872 0.004088 **
weekdaySaturday 1.353e-01 1.433e-02 9.444 < 2e-16 ***
weekdaySunday 1.312e-01 1.383e-02 9.489 < 2e-16 ***
weekdayThursday -4.865e-02 1.035e-02 -4.702 2.59e-06 ***
weekdayTuesday -5.870e-02 1.032e-02 -5.688 1.30e-08 ***
weekdayWednesday -4.643e-02 1.026e-02 -4.524 6.09e-06 ***
LDA_00 1.211e-01 2.313e-02 5.234 1.68e-07 ***

```

```

LDA_01 -1.136e-01 2.577e-02 -4.409 1.04e-05 ***
LDA_02 -8.478e-02 2.325e-02 -3.647 0.000266 ***
LDA_03 -1.069e-01 2.432e-02 -4.393 1.12e-05 ***
global_subjectivity 1.180e-01 3.821e-02 3.089 0.002011 **
global_rate_negative_words -9.250e-01 3.021e-01 -3.062 0.002200 **
min_positive_polarity -2.047e-01 4.678e-02 -4.376 1.21e-05 ***
max_positive_polarity -2.836e-02 1.598e-02 -1.775 0.075910 .
title_subjectivity 3.779e-02 1.102e-02 3.428 0.000609 ***
title_sentiment_polarity 3.151e-02 1.200e-02 2.625 0.008678 **
abs_title_subjectivity 6.367e-02 1.863e-02 3.418 0.000632 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 0.4364 on 21832 degrees of freedom
Multiple R-squared: 0.1446, Adjusted R-squared: 0.1426
F-statistic: 73.79 on 50 and 21832 DF, p-value: < 2.2e-16
summary(LogAICBIS)

##
Call:
lm(formula = shares ~ month + n_tokens_content + n_unique_tokens +
num_hrefs + num_imgs + average_token_length + num_keywords +
data_channel + kw_avg_avg + self_reference_avg_shares +
is_weekend + LDA_00 + LDA_01 + LDA_02 + LDA_03 + global_subjectivity +
global_rate_negative_words + avg_positive_polarity + abs_title_subjectivity +
abs_title_sentiment_polarity, data = popularityTrainBIS)
##
Residuals:
Min 1Q Median 3Q Max
-4.3411 -0.2832 -0.0612 0.2547 1.4394
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.868e+00 6.982e-02 84.047 < 2e-16 ***
month02 -2.443e-03 1.504e-02 -0.162 0.870946
month03 6.542e-02 1.526e-02 4.287 1.82e-05 ***
month04 5.052e-02 1.485e-02 3.401 0.000673 ***
month05 -5.123e-03 1.504e-02 -0.341 0.733436
month06 -9.423e-02 1.484e-02 -6.349 2.20e-10 ***
month07 -9.110e-02 1.474e-02 -6.182 6.45e-10 ***
month08 -6.399e-02 1.493e-02 -4.286 1.83e-05 ***
month09 -7.265e-02 1.496e-02 -4.858 1.19e-06 ***
month10 -6.957e-02 1.445e-02 -4.814 1.49e-06 ***
month11 -3.845e-02 1.510e-02 -2.546 0.010892 *
month12 -6.071e-02 1.546e-02 -3.928 8.59e-05 ***
n_tokens_content 1.823e-05 1.068e-05 1.706 0.087940 .
n_unique_tokens -1.572e-01 4.729e-02 -3.324 0.000890 ***
num_hrefs 1.643e-03 3.635e-04 4.519 6.25e-06 ***
num_imgs 7.107e-04 4.673e-04 1.521 0.128323
average_token_length -4.586e-02 1.192e-02 -3.847 0.000120 ***
num_keywords 6.356e-03 1.647e-03 3.860 0.000114 ***
data_channelEntertainment -5.795e-03 1.823e-02 -0.318 0.750533
data_channelLifestyle 4.965e-02 1.849e-02 2.686 0.007243 **
data_channelNone 1.235e-01 1.963e-02 6.289 3.25e-10 ***

```

```

data_channelSocial Media 2.215e-01 1.546e-02 14.333 < 2e-16 ***
data_channelTech 1.648e-01 1.677e-02 9.829 < 2e-16 ***
data_channelWorld 1.761e-02 1.793e-02 0.982 0.326193
kw_avg_avg 6.087e-05 3.086e-06 19.724 < 2e-16 ***
self_reference_avg_shares 6.411e-07 1.263e-07 5.075 3.92e-07 ***
is_weekend1 1.755e-01 9.030e-03 19.435 < 2e-16 ***
LDA_00 1.287e-01 2.325e-02 5.534 3.16e-08 ***
LDA_01 -1.017e-01 2.595e-02 -3.921 8.86e-05 ***
LDA_02 -1.101e-01 2.332e-02 -4.720 2.37e-06 ***
LDA_03 -6.941e-02 2.433e-02 -2.853 0.004337 **
global_subjectivity 1.733e-01 3.919e-02 4.421 9.87e-06 ***
global_rate_negative_words -9.158e-01 2.997e-01 -3.055 0.002250 **
avg_positive_polarity -9.561e-02 3.922e-02 -2.438 0.014781 *
abs_title_subjectivity 4.378e-02 1.749e-02 2.503 0.012321 *
abs_title_sentiment_polarity 5.626e-02 1.495e-02 3.762 0.000169 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 0.4408 on 21844 degrees of freedom
Multiple R-squared: 0.1265, Adjusted R-squared: 0.1251
F-statistic: 90.37 on 35 and 21844 DF, p-value: < 2.2e-16

```

Zastosowanie AIC daje nam podobne efekty jak BIC, zmniejszamy ilość zmiennych, lecz niedźżum kosztem  $R^2$ .

## Regresja grzbietowa & Lasso

Kolejnymi metodami możliwymi do zaaplikowania są: Regresja grzbietowa oraz LASSO. Są one do siebie zbliżone. Nie przyniosły one porządkanych wyników dla tego zrezygnowaliśmy z ich prezentacji.

## Walidacja & Testowanie

### Walidacja

Posiadamy kilka skonstruowanych modeli, teraz porównamy je na zbiorze walidacyjnym, sprawdzimy jak się zachowują i wybierzemy ten najlepszy (oczywiście najpierw musimy odpowiednio zmodyfikować nasz zbiór):

```

summary(popularityLog)$r.squared

[1] 0.1447832
1-var(queryPopularityLog)/var(popularityQuery$shares)

[1] 0.1298482
summary(popularityLogBIS)$r.squared

[1] 0.1266817
1-var(queryPopularityLogBIS)/var(popularityQuery$shares)

[1] 0.1111866
summary(LogBIC)$r.squared

[1] 0.1420315

```

```

1-var(queryPopularityLogBIC)/var(popularityQuery$shares)

[1] 0.1254836
summary(LogBICBIS)$r.squared

[1] 0.1249669
1-var(queryPopularityLogBICBIS)/var(popularityQuery$shares)

[1] 0.1094853
summary(LogAIC)$r.squared

[1] 0.1445574
1-var(queryPopularityLogAIC)/var(popularityQuery$shares)

[1] 0.1297525
summary(LogAICBIS)$r.squared

[1] 0.1264849
1-var(queryPopularityLogAICBIS)/var(popularityQuery$shares)

[1] 0.1111507

```

Po porównaniu modeli, okazuje się, że najlepsze dopasowanie wykacuje model **LogAIC**. To na nim sprawdzono jak zachowuje się ostatni zbiór testowy.

## Testowanie

Sprawdźmy:

```

queryPopularityLogAIC=popularityTest$shares-predict(LogAIC, popularityTest)
1-var(queryPopularityLogAIC)/var(popularityTest$shares)

[1] 0.1417775

```

Wszystko wygląda w porządku. Oczywiście o ile dopasowanie na poziomie 14% można uznać za zadowalające.

## Podsumowanie/posłowie

Udało nam się zbudować model, który przy użyciu stosunkowo małej ilości zmiennych potrafi przewidzieć liczbę udostępnień artykułu z dokładnością na poziomie 14%. Jednak za tym wszystkim kryje się haczyk...

Nasze dane były totally pomieszcane, nie ujawniono tego wcześniej, ale podczas analizy znaleziono błąd. Następnie ręcznie, metodą bisekcji, znaleziono pierwszy błędny wiersz w danych, jest to linijka: **8011** oraz artykuł Is This Tumblr's New Photo-Rich Redesign? od tego miejsca, ilość **share'ów** jest przesunięta względem pozostałych danych, z czasem różnica wierszy rośnie do kilku, poprawienie rekordów ręcznie wymagałoby cierpliwości i byłoby niesamowicie czasochłonne. Warto z tego wziąć nauczkę na przyszłość, żeby bardzo dokładnie przyjrzeć się danym, także poprzez np. losowe sprawdzenie pewnych informacji.