Uniwersytet Warszawski

Wydział Nauk Ekonomicznych

Łukasz Madej

Co wpływa na popularność filmów? Analiza czynników wpływających na ilość wyświetleń treści wideo na platformie YouTube

Praca przygotowana na
zajęcia z ekonometrii
pod kierunkiem mgr. Rafała Walaska

Warszawa Styczeń 2025

Spis treści

1. Wstęp

W erze cyfryzacji obserwujemy dynamiczny wzrost wielu internetowych platform, takich jak YouTube, której funkcjonowaniu się w niniejszej pracy przyjrzymy. Treści wideo online stały się jednym z głównych źródeł rozrywki, informacji oraz edukacji dla milionów użytkowników na całym świecie. Pandemia koronawirusa dodatkowo nasiliła ten trend - wraz z przymusową izolacją światowe społeczeństwo zaczęło spędzać jeszcze więcej czasu online, co utrzymuje się aż po dzisiejszy dzień. Jednak to, co decyduje o sukcesie danego filmu nie zawsze jest oczywiste. Konkurencja wśród twórców stale rośnie, co jest wynikiem zarówno proliferacji kanałów o znacznym zapleczu finansowym, jak i co raz bardziej zaawansowanych algorytmów promujących treści.

Wzrost konkurencji sprawia, że dzisiejsi twórcy są w potencjalnie znacznie trudniejszej sytuacji niż chociażby piętnaście lat temu, kiedy platforma nie była jeszcze aż tak rozwinięta. Aby przyciągnąć i zbudować widownię, która umożliwi im osiągać zyski materialne na podstawie monetyzacji swoich filmów, twórcy muszą ostrożnie zastanowić się nad zagadnieniem tego, jak działa algorytm promujący ich materiały.

W tej pracy za pomocą modelu regresji liniowej przyjrzymy się, które czynniki i w jaki sposób determinują liczbę wyświetleń danego filmu. Główną hipotezą, którą zweryfikujemy jest pozytywny wpływ komentarzy na liczbę wyświetleń. Jest to jedna z głównych form interakcji ze strony widowni, na której zazwyczaj zależy twórcom. Może być to związane z efektem sieciowym – większa liczba komentarzy potencjalnie sygnalizuje zwiększone zainteresowanie materiałem, co może zachęcić kolejne osoby do obejrzenia filmu i zwiększać jego faworyzację przez algorytmy. Następnie zweryfikujemy, czy filmy opublikowane przed rokiem 2020 są bardziej popularne od tych opublikowanych po roku 2020. Jest to ciekawy temat, ze względu na dwa przeciwstawne czynniki. Z jednej strony starsze filmy miały więcej czasu na generowanie wyświetleń, co daje im przewagę czasową. Z drugiej strony filmy opublikowane w trakcie i po pandemii, znajdowały się w korzystniejszym środowisku pod względem liczby użytkowników platformy i potencjalnie większej widowni w sytuacji nadmiernej popularności filmu krótko po opublikowaniu. Ostatnią hipotezą, którą zweryfikujemy jest pozytywny wpływ długości filmu na wyświetlenia. Efektem może tutaj być percepcja wartości postrzeganej – dłuższe materiały mogą być interpretowane jako bardziej wartościowe, oferujące głębszą analizę lub bardziej rozbudowaną rozrywkę.

2. Analiza literatury w kontekście modelu

Popularność filmu zależy od wielu różnorodnych czynników. To zagadnienie budzi zainteresowanie nie tylko twórców treści, ale także specjalistów od marketingu i projektantów platform. Istotną rolę odgrywają czynniki trudne do zmierzenia, takie jak to, czy film jest wysokiej jakości, interesujący, informacyjny lub budzący emocje. Film o takich cechach znacznie łatwiej zaangażuje widownię, która ma znaczący wpływ na promowanie filmu, co zbadał Jiaqi Shen za pomocą regresji liniowej. Najważniejszą interakcją od strony widza są właśnie komentarze, a następnie polubienia. Za pomocą interakcji tych dwóch zmiennych autor zauważył, że najlepsze filmy zawierają wysoką wartość tych dwóch czynników jednocześnie. Zaskakującym wnioskiem jest to, że liczba „łapek w dół" również ma pozytywny wpływ na popularność filmu, co można tłumaczyć wcześniej wspomnianym wywoływaniem emocji. To, co nas interesuje w kontekście modelu to znaczący wpływ metryk związanych z interakcją od strony widowni. (Shen, 2024).

Wyświetlenia nie zależą jednak tylko od samych interakcji i jakości filmu. Autorzy Raphaela M. Velho, Amanda M. F. Mendes, Caio L. N. Azevedo zbadali 441 filmów z kanałów należących do projektu ScienceVlogs Brasil, który zresza twórców treści naukowych w języku portugalskim. Zauważyli również, jak autor poprzedniego artykułu, że komentarze mają znaczny wpływ na popularność. Ze względu na problemy z kolinearnością autorzy wykorzystali liczbę polubień jako jedyną zmienną związaną z zaangażowaniem widowni w swoim modelu ekonometrycznym. Wzięli również pod uwagę dość nietuzinkowe czynniki, takie jak produktywność twórcy, która jest mierzona przez częstotliwość wrzucania filmów, wiek filmów oraz typ filmów. Co ciekawe, wiek filmu wskazywał na negatywny efekt na zmienną objaśnianą, jednak to może być wytłumaczone naturą wybranych filmów – starsze materiały były opublikowane w niekorzystnych formatach przez twórców o niskiej produktywności, zaś nowe przez twórców o wyższej produktywności, co mogło mieć negatywny wpływ na algorytm. Warto zwrócić uwagę na długość filmu, która w tym badaniu wykazała pozytywny wpływ na oglądalność (Velho, Mendes, Azevedo, 2020).

Ciekawe podejście do tematu obrali Dustin J. Welbourne i Will J. Grant. W swoim badaniu przeanalizowali 390 filmów z 39 różnych kanałów związanych z nauką - zarówno profesjonalnych, które zawierały całe zespoły poświęcone produkcji treści, jak i tworzonych przez zwykłych użytkowników-amatorów. Ku mojemu zdziwieniu, to właśnie ta druga grupa okazała się być bardziej efektywna w zdobywaniu większej oglądalności. Istotnymi czynnikami były również szybkość narracji oraz powracający narrator, co autorzy starali się wytłumaczyć

przywiązaniem widowni do ulubionych twórców. W moim modelu niestety nie będę mógł zweryfikować istotności tych zagadnień ze względu na ograniczenie moich danych, jednak powinniśmy przyjrzeć się innej kwestii. Odwrotnie niż w poprzednim artykule, autorzy przekonują nas o znikomym wpływie długości filmu na liczbę wyświetleń. Dwa sprzeczne sygnały dają nam kolejny powód, że warto jest tą zmienną zbadać. (Welbnourne, Grant 2015).

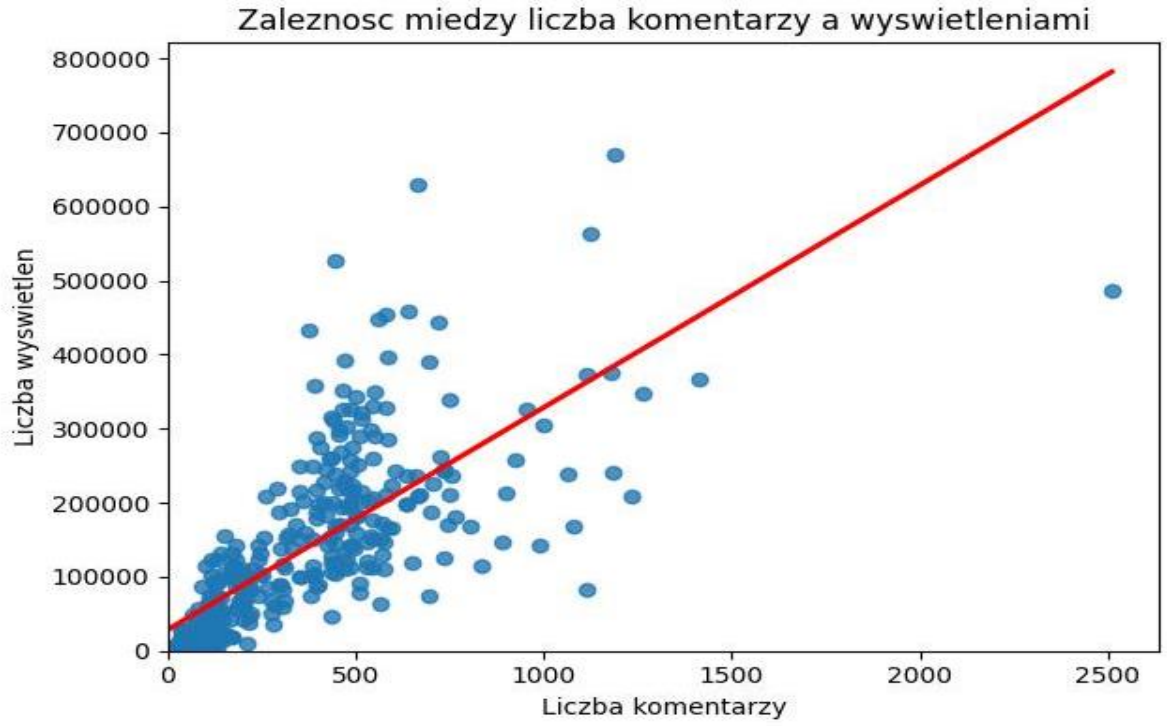
Najbardziej obszernym, bo analizującym aż 37 milionów filmów, jest badanie, gdzie popularność materiałów zbadali autorzy: Gloria Chatzopoulou, Cheng Sheng i Michalis Faloutsos. W pracy skupili się na czterech metrykach popularności: liczbie komentarzy, ocen, "dodanych do ulubionych" oraz średniej ocenie. Jednym z głównych odkryć autorów jest "magiczna liczba" algorytmu - na każde 400 wyświetleń przypada jeden komentarz, jedna ocena i jedno dodanie filmu do ulubionych. Autorzy tworząc model regresji liniowej zauważyli, że te trzy wymienione czynniki mają znaczący wpływ na popularność, zaś średnia ocena okazała się być nieistotna. Jest to kolejna przesłanka, informująca nas o istotności zmiennych związanych z zaangażowaniem widowni. Zaobserwowano w tym badaniu również znaczenie podtrzymywania wysokiej oglądalności filmu w krótkim okresie po publikacji - wtedy jest największa szansa na to, że film się osiągnie sukces. Podobnie jak w pierwszym artykule, autorzy podkreślają znaczenie optymalizacji treści w celu maksymalizacji interakcji użytkowników, która ma kluczową rolę w generowaniu wyświetleń. Zwrócono również uwagę na sieć powiązanych filmów - algorytm zdaje się faworyzować podobne treści pod względem tytułu, słów kluczowych oraz opisu - co również wydaje się być istotną informacją dla twórców, ale niestety w modelu nie mamy jak tego zbadać. (Chatzopoulou, Sheng, Faloutsos, 2010).

Ostatni artykuł zgłębia aspekt wcześniej wspomnianego mechanizmu polecania filmów, który został powiązany z wartością *click-through rate* CTR, a zatem wskaźnikiem oceniającym to, jak często film jest otwierany po wstępnej rekomendacji. Badanie szczegółowo omawia CTR jako ważną determinantę efektywności powiązanych filmów. Autorzy Renjie Zhou, Samamon Khemmarat, Lixin Gao, Jian Wan, Jilin Zhang, Yuyu Yin i Jun Yu przeanalizowali 100 najpopularniejszych filmów w historii platformy, aby zrozumieć jak mechanizmy platformy przekładają się na oglądalność filmów. To właśnie rekomendacje i wyszukiwanie są najbardziej trwałymi źródłami ruchu, które generują stabilne tempo wzrostu liczby wyświetleń w czasie. Filmy, które pojawiają sięna szczycie list rekomendacji lub na stronie głównej, charakteryzują się najwyższym wskaźnikiem CTR. To sugeruje, że popularniejsze filmy będą zawierały większą wartość owego wskaźnika, co jest przesłanką, że warto mu się przyjrzeć podczas analizy ekonometrycznej. (Zhou, Khemmarat, Gao, Wan, Zhang, Yin, Yu, 2016).

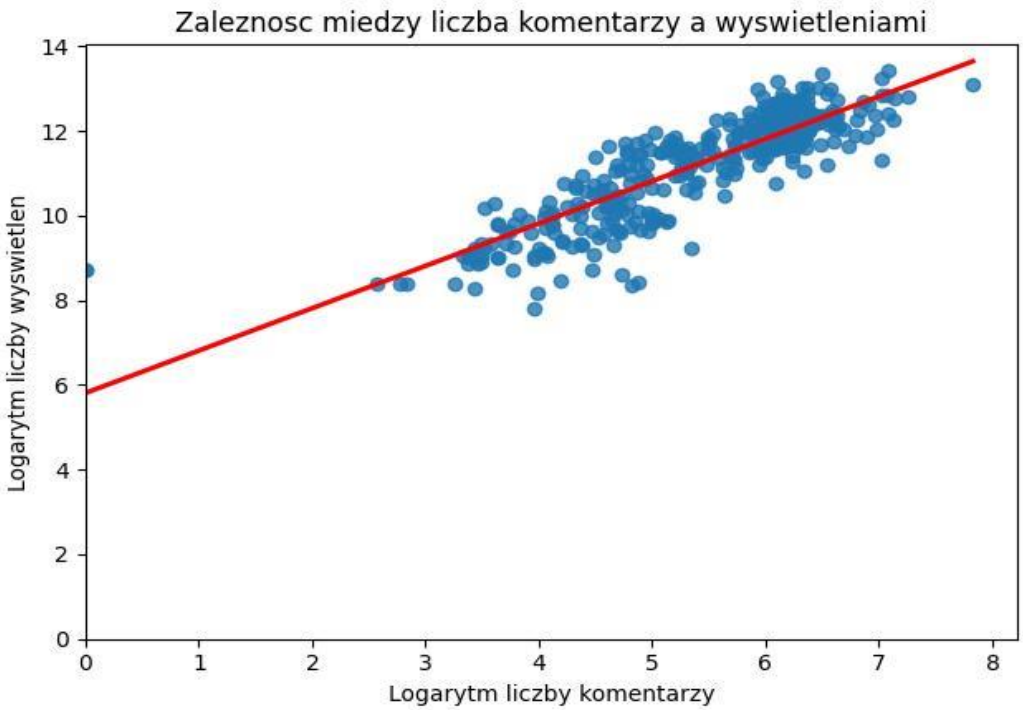
3. Analiza wstępna

Dane, na których stworzyłem mój model pochodzą z platformy Kaggle, ze zbioru danych pt. “YouTube Channel Performance Analytics”. W zbiorze zostały przedstawione 364 filmy wraz z informacjami na ich temat, pozyskane za pomocą narzędzia analitycznego platformy YouTube. Zakres czasowy jest bardzo obszerny, gdyż obejmuje aż 8 lat - najstarsze filmy zostały opublikowane w roku 2016, najnowsze zaś pod koniec 2024.

Postaramy się przyjrzeć kluczowym czynnikom, które mogą mieć wpływ na liczbę wyświetleń. Pierwszą i główną zmienną, której się przeanalizujemy, jest liczba komentarzy. W trzech przedstawionych wcześniej artykułach ogromna waga przywiązywana była właśnie do czynników odpowiadających za zaangażowanie widowni.



Powyższy wykres pokazuje zależność między liczbą komentarzy, a wyświetleniami. Jest to jeden z priorytetowych czynników, zatem ważne było wybranie dogodnej formy naszych zmiennych – zauważalna jest mocna koncentracja punktów dla niskich wartości komentarzy oraz wyświetleń, dlatego zdecydowałem się na transformację logarytmiczną.



Możemy zaobserwować, że nasza relacja coraz bardziej przypomina zależność liniową. Dzięki powyższej transformacji udało się wyeliminować problem mocno odstających reszt.

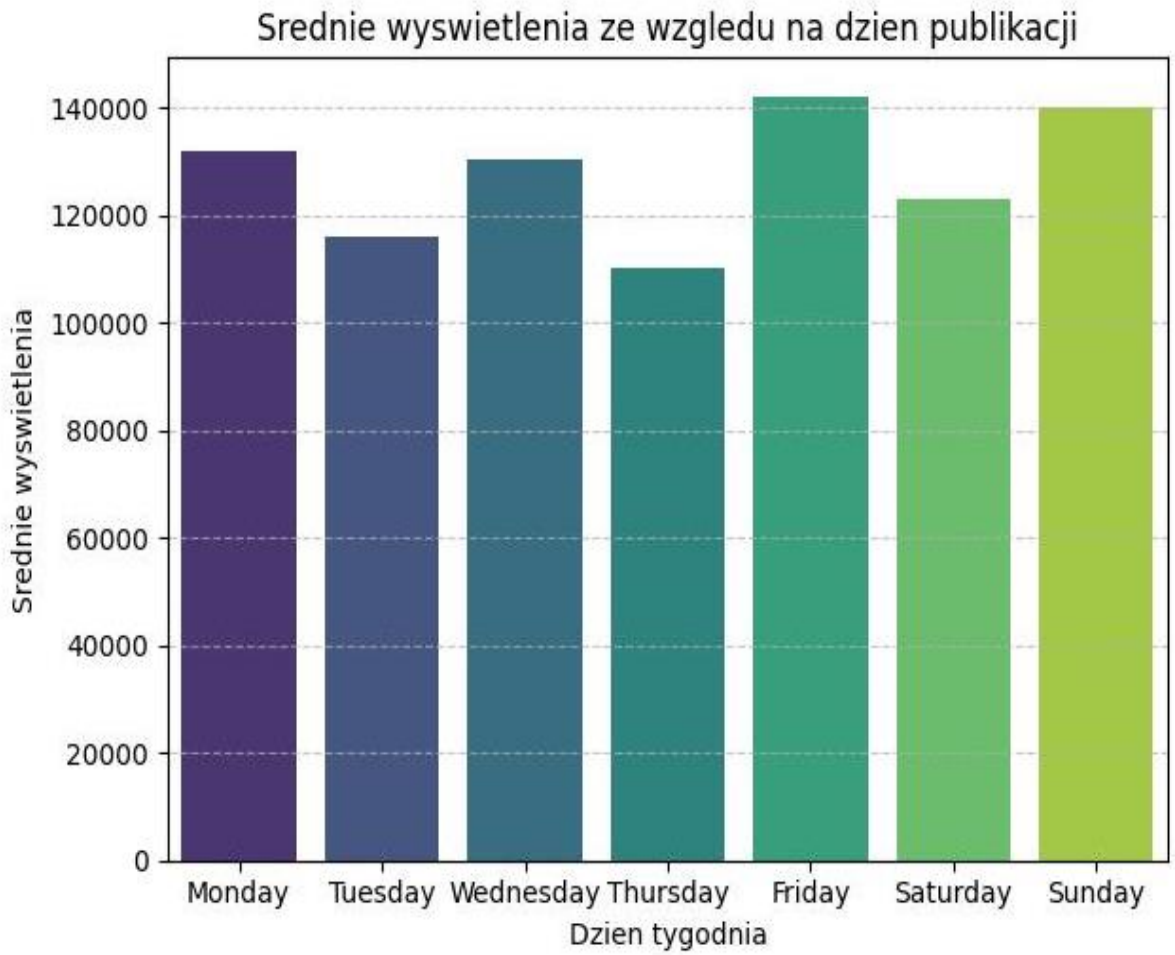
Długość filmu określona w sekundach również zostanie uwzględniona w moim modelu. Długość filmu może być związana z postrzeganą wartością materiału przez widownię, a jej wpływ na popularność, jak wynika z literatury, pozostaje niejednoznaczny, co czyni ją interesującym czynnikiem do analizy.

Kolejną analizowaną zmienną będzie procentowy współczynnik CTR. Ta determinanta pozwala nam poniekąd zbadać wpływ czynników, które trudno zmierzyć, takich jak optymalizacja tytułu, miniatury filmu, a także faworyzację algorytmu w promowaniu treści.

Ostatnim analizowanym czynnikiem w postaci ciągłej jest estymowany przychód w dolarach pochodzący z filmu. Nie zostało to opisane w literaturze, aczkolwiek spotkałem się z tezą, że filmy, w których pojawia się więcej reklam powiązane są z większą faworyzacją algorytmu, co wydaje się być dość logiczne, biorąc pod uwagę zysk platformy w takiej sytuacji.

Jako zmienną kategoryzującą wykorzystałem dzień tygodnia, w którym został opublikowany film. Z literatury wynika, że na popularność filmu znaczny wpływ ma jego początkowa wydajność. Publikacja filmu w trakcie tygodnia może mieć negatywny wpływ na jego wczesną popularność, niż w sytuacji opublikowania treści w weekend. Wtedy potencjalna

widownia dysponuje większą ilością wolnego czasu, który może przełożyć się na większą oglądalność.



Histogram wskazuje na potencjalne zróżnicowanie efektywności filmów w zależności od tygodnia ich publikacji. Zastanawiająca jest sobota, ponieważ można by oczekiwać, że jej wartość będzie równie wysoka jak w pozostałe dni weekendu.

Ostatnim czynnikiem, który stworzyłem własnoręcznie i postanowiłem dodać do modelu jest zmienna zero-jedynkowa *past2020* – która jak możemy się domyślić, determinuje to, czy film został dodany przed lub po roku 2020. Z jednej strony starsze filmy mogą korzystać z dłuższego czasu na gromadzenie wyświetleń, zaś nowsze mogą czerpać korzyści z efektu większej liczby użytkowników podczas publikacji oraz efektywniejszych rozwiązań twórców.

	video duration	comments	shares	dislikes	likes	views	estimated revenue	ctr
count	364	364	364	364	364	364	364	364
mean	664.23	333.84	252.95	123.96	5526.73	128800.10	8.85	7.91
std	330.64	291.93	363.01	128.31	4465.21	118209.84	13.41	2.90
min	9	0	1	2	121	2461	0	0.62
med	613	262.5	149	85.5	5172	101950.50	4.28	8.39
max	2311	2510	4190	818	27222	670990	103.11	27.66

Powyższa tabela przedstawia podstawowe statystyki opisowe dla analizowanych zmiennych liczbowych. Dane obejmują łącznie 364 obserwacje. Najmniej popularny film uzyskał 2461 wyświetleń, a najbardziej - 670990. Średnio każdy film wygenerował ponad sto tysięcy wyświetleń, co można uznać za bardzo dobry wynik. Zaskakująca jest wartość estymowanego przychodu, gdyż jej średnia wynosi zaledwie 8,85 USD na film. Powodem może być widownia z mniej zamożnych części świata – interakcje z reklamami od widzów z krajów rozwiniętych są bowiem wyżej wyceniane w wynagrodzeniu twórcy od reklamodawcy.

Ostatecznie, oprócz wyświetleń i komentarzy, zlogarytmowałem również zmienne dotyczące estymowanego przychodu oraz długości filmu. Przekształcenie to pozwoliło uzyskać bardziej liniowe zależności, zachowując jednocześnie sensowność ekonomicznej interpretacji tych czynników. Dodałem również wartość jeden do wartości kolumny z komentarzami oraz estymowanym przychodem, ze względu na występowanie wielu wartości zerowych, a dla przychodu wartości w przedziale od 0 do 1, co mogło zachwiać moim modelem po transformacji logarytmicznej.

Model regresji liniowej, estymowany metodą najmniejszych kwadratów, przyjmuje następującą postać:

$$\begin{aligned} \ln views = {} & \beta_0 + \beta_1 \ln comments + \beta_2 \ln estimated\_revenue + \beta_3 \ln video\_duration + \\ & \beta_4 vid\_thumbnail\_ctr\_percent + \beta_5 past2020 + \beta_6 Monday + \beta_7 Tuesday + \\ & \beta_8 Wednesday + \beta_9 Thursday + \beta_{10} Friday + \beta_{11} Saturday + \varepsilon_i \end{aligned}$$

Gdzie zmienną objaśnianą jest logarytm liczby wyświetleń (y = ln views). Zmienne objaśniające, których wybór szczegółowo uzasadniłem w kolejnym rozdziale, to:

1. x_1 = **lncomments** – logarytm naturalny liczby komentarzy, zmienna ciągła,
2. x_2 = **lnestimated_revenue** – logarytm naturalny estymowanego przychodu w dolarach, zmienna ciągła,
3. x_3 = **lnvideo_duration** – logarytm naturalny długości trwania filmu w sekundach, zmienna ciągła,
4. x_4 = **vid_thumbnail_ctr_percent** – wskaźnik „klikalności" filmu w sytuacji gdy zostaje on zaproponowany na stronie, zmienna ciągła przyjmująca wartości od 0 do 1,
5. x_5 = **past2020** - zmienna zero-jedynkowa przyjmująca wartość 1 jeśli film został opublikowany po roku 2020, w przeciwnym wypadku przyjmuje 0,
6. x_6 = **Monday** - zmienna zero-jedynkowa, przyjmująca wartość jeden jeśli film został opublikowany w poniedziałek; poziomem bazowym dla wszystkich zmiennych dotyczących dni tygodnia jest niedziela,
7. x_7 = **Tuesday** - zmienna zero-jedynkowa, przyjmująca wartość jeden jeśli film został opublikowany we wtorek,
8. x_8 = **Wednesday** - zmienna zero-jedynkowa, przyjmująca wartość jeden jeśli film został opublikowany w środę,
9. x_9 = **Thursday** - zmienna zero-jedynkowa, przyjmująca wartość jeden jeśli film został opublikowany w czwartek,
10. x_{10} = **Friday** - zmienna zero-jedynkowa, przyjmująca wartość jeden jeśli film został opublikowany w piątek,
11. x_{11} = **Saturday** - zmienna zero-jedynkowa, przyjmująca wartość jeden jeśli film został opublikowany w sobotę.

4. Analiza ekonometryczna

Na początku poruszę analizę korelacji zmiennych metodą Spearmana. Jak wcześniej się dowiedzieliśmy, bardzo ważne są czynniki powiązane z zaangażowaniem widza - komentarze, polubienia, "łapki w dół", oraz udostępnienia. Każda z nich wysoko koreluje z wyświetleniami (>0.77), jednak są również mocno skorelowane między sobą. Najmniejsza wartość wynosi 0.64, jednak dla większości z nich wynosi ona znacznie więcej i oscyluje w okolicach 0.8. Aby uniknąć wielokrotnego wnoszenia tej samej informacji do modelu, postanowiłem wybrać tylko jedną z nich, mianowicie komentarze, które korelują z wyświetleniami na poziomie 0.86. Następnie, po ostrożnym wyselekcjonowaniu wybrałem również zmienne estymowanego przychodu oraz wskaźnik CTR, które w sposób dostateczny (kolejno 0.23 i 0.43) korelują ze zmienną objaśnianą, ale nie z komentarzami, tak jak zmienne związane z zaangażowaniem.

Długość filmu natomiast koreluje w sposób negatywny z wyświetleniami (-0.024), jednak mimo wszystko uwzględniłem tą zmienną modelu – jej istotność zostanie określona na późniejszym etapie analizy.

Ostatecznie zdecydowałem się nie uwzględniać nieliniowości oraz interakcji, ponieważ analiza danych nie dostarczyła podstaw do takiego podejścia. Podczas analizy wykresów zależności wyświetleń od zmiennych objaśniających w formie kwadratowej nie zauważyłem, aby istniała taka zależność. W odniesieniu do interakcji - analizując wykresy zależności między zmiennymi oraz ich rozkładów rozdzielając je na dwie grupy, w zależności od wartości zmiennej *past2020*, nie zauważyłem istotnych różnic pomiędzy nimi. Kolejnym powodem jest fakt, że chcemy zbadać wpływ tej zmiennej na wyświetlenia. Status istotności zmiennej, która weszłaby z nią w interakcję mógłby przesądzić o istotności lub nieistotności *past2020*, co było kolejną przesłanką kształtującą moją decyzję.

Na wydruku na następnej stronie znajdują się cztery modele, dla których zastosowałem metodę od ogółu do szczegółu, w której kolejno sprawdzałem łączną nieistotność zmiennych nieistotnych o największym *p-value*. Aby wydruk się nie zniekształcił w poniższym raporcie, pozwoliłem sobie na skrócenie nazw zmiennych kategorycznych odnośnie dnia tygodnia, jednak nadal poziomem bazowym jest niedziela.

Przyjąłem poziom istotności równy 0,05. W przejściu z modelu pierwszego do drugiego usunąłem zmienne kategoryczne dotyczące dni tygodnia - każda z nich miała wartość *p-value* większą niż 0,05, a dwie z nich miały jej największą wartość ze wszystkich zmiennych w modelu - okazały się być łącznie nieistotne.

Przechodząc z modelu drugiego do trzeciego, postanowiłem usunąć zmienną zlogarytmowanej długości filmu – spośród pozostawionych zmiennych posiadała najwyższe *p-value* = 0,78 i jest ona łącznie nieistotna wraz ze zmiennymi odpowiadającymi za dzień tygodnia.

W ostatnim kroku, przechodząc z modelu trzeciego do czwartego usunąłem zmienną zero-jedynkową *past2020* - jej *p-value* wynosiło 0,22 i była ona łącznie nieistotna z poprzednio wyrzuconymi zmiennymi – sprawdzając hipotezę o łącznej nieistotności wszystkich wyrzuconych zmiennych otrzymałem *p-value* = 0,61, co oznacza brak podstaw do odrzucenia hipotezy o łącznej nieistotności tych zmiennych.

Aby zweryfikować łączną nieistotność wszystkich zmiennych, przyjrzyjmy się bazowemu, pierwszemu modelowi. Statystyka F o wartości 144,074 i jej p-value = 0 neguje tę tezę – zmienne objaśniające zawarte w modelu są łącznie istotne. Warto jest się również przyjrzeć istotności pojedynczej, najważniejszej zmiennej w moim modelu – logarytmowi z liczby komentarzy. Na podstawie testu t-Studenta, otrzymamy wartość p-value = 0, co wskazuje na istotność tej zmiennej.

	(1)	(2)	(3)	(4)
Intercept	5.946***	5.928***	5.871***	5.790***
	(0.279)	(0.267)	(0.161)	(0.146)
np.log(comments)	0.818***	0.815***	0.812***	0.827***
	(0.035)	(0.034)	(0.032)	(0.030)
np.log(estimated_revenue)	0.100***	0.099***	0.098***	0.090***
	(0.018)	(0.017)	(0.017)	(0.015)
np.log(video_duration)	-0.014	-0.011		
	(0.043)	(0.043)		
vid_thumbnail_ctr_percent	0.113***	0.112***	0.112***	0.110***
	(0.011)	(0.011)	(0.011)	(0.011)
C(past2020)[T.True]	-0.090	-0.080	-0.085	
	(0.075)	(0.072)	(0.070)	
C(week_day)[T.Monday]	-0.173			
	(0.105)			
C(week_day)[T.Tuesday]	0.018			
	(0.098)			
C(week_day)[T.Wednesday]	-0.045			
	(0.099)			
C(week_day)[T.Thursday]	-0.054			
	(0.114)			
C(week_day)[T.Friday]	-0.011			
	(0.100)			
C(week_day)T.Saturday]	0.035			
	(0.104)			
Observations	364	364	364	364
R^2	0.818	0.816	0.816	0.815
Adjusted R^2	0.813	0.813	0.814	0.813
Residual Std. Error	0.537	0.536	0.536	0.536
F Statistic	144.074***	317.120***	397.409***	528.728***

Note: *p<0.1; **p<0.05; ***p<0.01

Dependent variable: np.log(views)

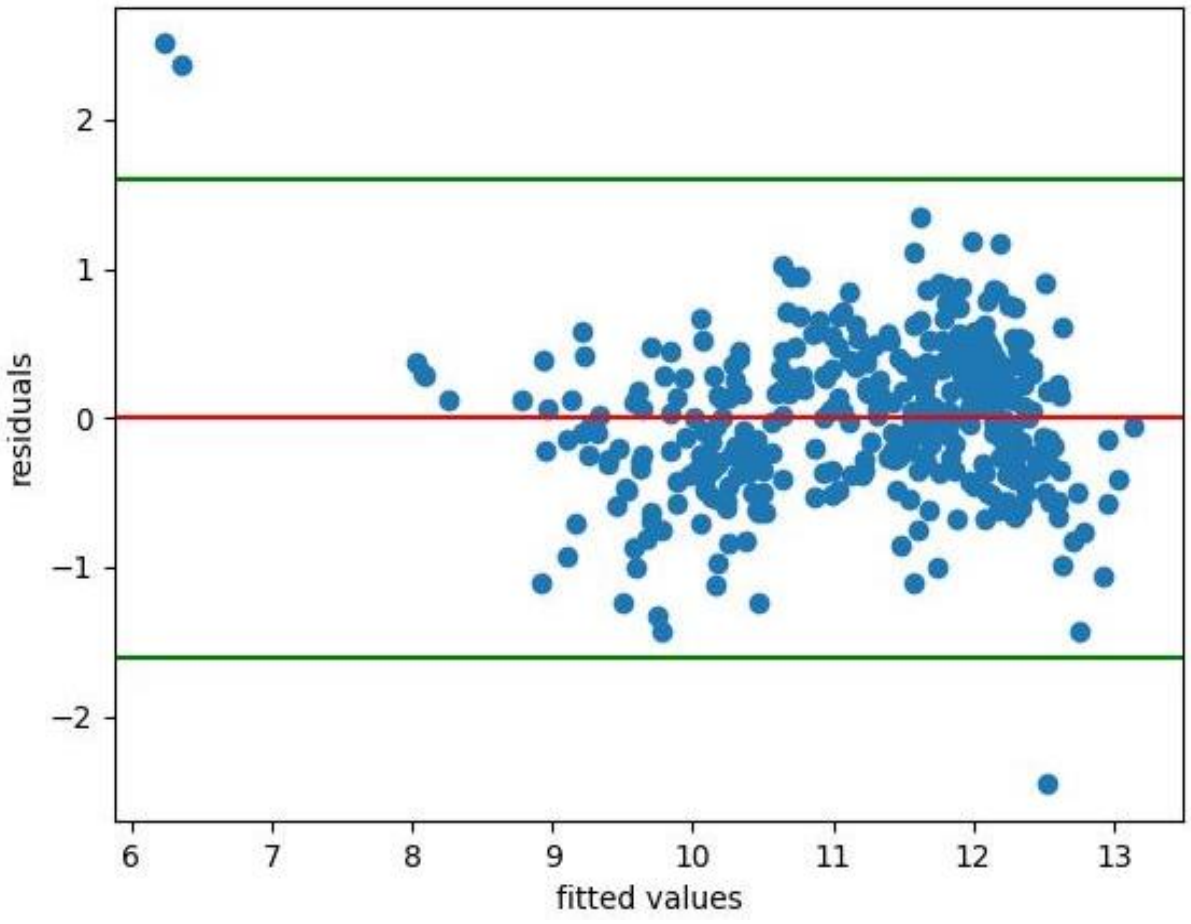
Przechodząc ostatecznie do modelu czwartego pozostaliśmy z modelem, który zawiera jedynie zmiennie istotne statystycznie. Skorygowany współczynnik determinacji we wszystkich modelach pozostaje na stabilnym poziomie, jednak widzimy że w modelu trzecim jest on największy. Różnica jest jednak na tyle znikoma, że w ostatecznej decyzji uznałem

model czwarty za najlepszy, ze względu na istotność wszystkich zmiennych objaśniających. Na podstawie wartości współczynnika determinacji możemy stwierdzić, że modelem udało się nam wyjaśnić 81,5% zmienności zmiennej objaśnianej.

Z modelu wynika, że wzrost liczby komentarzy i estymowanego przychodu w dolarach o 1% przekłada się odpowiednio na wzrost wyświetleń o 0,827% i 0,090%. Z kolei wzrost wskaźnika CTR o jeden punkt procentowy skutkuje wzrostem wyświetleń o 11%. Nie możemy jednak swobodnie interpretować ekonomicznie tych parametrów, gdyż można je podważyć ze względu na niespełnienie testu na poprawność przyjętej formy funkcyjnej, który opiszę w następnym rozdziale.

5. Analiza założeń KMRL

W poniższym rozdziale sprawdzę, czy mój model spełnia założenia klasycznego modelu regresji liniowej (KMRL). Na początku przyjrzę się założeniu o homoskedastyczności reszt, przeprowadzając test Breuscha-Pagana oraz test White'a. Zanim jednak przejdę do formalnych testów, przeanalizujemy wykres przedstawiający zależność reszt od wartości dopasowanych.



Na powyższym wykresie widać, że wraz ze wzrostem wartości dopasowanych odległość reszt od linii referencyjnej y = 0 zwiększa się, co może sugerować występowanie problemu heteroskedastyczności reszt w modelu.

Nazwa testu	Testowane założenie KMRL	Hipoteza zerowa	Statystyka testowa	p-value	Decyzja weryfikacyjna
Test Breuscha-Pagana	Wariancja składnika losowego jest identyczna dla wszystkich obserwacji	Homoskedastyczność reszt	38.48	0	Odrzucamy H0 o homoskedastyczności reszt na rzecz H1 o heteroskedastyczności reszt.

Nazwa testu	Testowane założenie KMRL	Hipoteza zerowa	Statystyka testowa	p-value	Decyzja weryfikacyjna
Test White'a	Wariancja składnika losowego jest identyczna dla wszystkich obserwacji	Homoskedastyczność reszt	240.89	0	Odrzucamy H0 o homoskedastyczności reszt na rzecz H1 o heteroskedastyczności reszt.

Dwa powyższe testy wskazują na występowanie heteroskedastyczności reszt w modelu, co skutkuje obciążeniem macierzy wariancji-kowariancji estymatora MNK. W efekcie mogą wystąpić błędne wyniki wnioskowania statystycznego przy wykorzystaniu standardowych statystyk testowych.

Aby rozwiązać ten problem, postanowiłem przeanalizować model wyjaśniający reszty do kwadratu, z tymi samymi zmiennymi objaśniającymi. Istotność zmiennych w takim modelu implikuje, że to one mogą powodować problem heteroskedastyczności. Zmienną istotną statystycznie okazał się być logarytm liczby komentarzy. Jest to o tyle problematyczne, że w teorii zlogarytmowanie zmiennej powinno pomóc z heteroskedastycznością, jednak nie możemy tego zrobić – zmienna jest już zlogarytmowana.

Co ciekawe, po cofnięciu transformacji logarytmicznej zmienna przestała być istotna statystycznie w modelu wyjaśniającym reszty podniesione do kwadratu, aczkolwiek stała pozostała statystycznie istotna. Model bez transformacji logarytmicznej komentarzy nadal nie spełniał założeń o homoskedastyczności, dlatego ostatecznie zdecydowałem się na zastosowanie macierzy odpornej White'a, która nie zniweluje tego problemu, aczkolwiek uodporni nasz model.

	Dependent variable: np.log(views)	
	(1)	(2)
Intercept	5.790***	5.790***
	(0.146)	(0.266)
np.log(comments)	0.827***	0.827***
	(0.030)	(0.065)
np.log(estimated_revenue)	0.090***	0.090***
	(0.015)	(0.015)
vid_thumbnail_ctr_percent	0.110***	0.110***
	(0.011)	(0.025)
Observations	364	364
R^2	0.815	0.815
Adjusted R^2	0.813	0.813
Residual Std. Error	0.536	0.536
F Statistic	528.728***	183.806***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Powyższy wydruk przedstawia model czwarty przed (1) i po (2) zastosowaniu macierzy odpornej White'a. Zauważalny jest wzrost wartości błędów standardowych dla wszystkich zmiennych oprócz logarytmu estymowanego przychodu. Statystyka F całego modelu również uległa obniżeniu. Wszystkie zmienne pozostały statystycznie istotne.

Następnie przeprowadzimy test Ramseya RESET, który zweryfikuje, czy nasz model ma poprawną, liniową formę funkcyjną.

Nazwa testu	Testowane założenie KMRL	Hipoteza zerowa	Statystyka testowa	p-value	Decyzja weryfikacyjna
Test RESET	Poprawna forma funkcyjna modelu	Model jest liniowy	122.75	0	Odrzucamy H0 o liniowości modelu na rzecz H1 o nieliniowości modelu

Model nie spełnia tego założenia, co sugeruje, że estymator MNK jest obciążony. Nie jesteśmy w stanie udowodnić nieobciążoności i efektywności estymatora. W związku z tym,

model nie powinien być używany do wyciągania ekonomicznych wniosków, co podkreśliłem przy interpretacji oszacowań. Aby rozwiązać problem nieliniowości modelu, tworzyłem kolejne modele, dodając następne potęgi zmiennych oraz interakcje, jednak po sześciu próbach problem nadal się utrzymywał. Nieliniowość została zażegnana dopiero w modelu wielomianowym do trzeciej potęgi.

Następnie zbadałem stabilność parametrów w podpróbkach. Wykorzystałem zmienną *past2020* do podzielenia danych na dwie grupy – filmów opublikowanych przed i po 2020 rokiem.

Nazwa testu	Testowane założenie KMRL	Hipoteza zerowa	Statystyka testowa	p-value	Decyzja weryfikacyjna
Test Chowa	Jednorodność parametrów w różnych podgrupach	Parametry w podpróbkach są stabilne	3.20	0.013	Odrzucamy H0 o stabilności parametrów w podpróbkach na rzecz niestabilności parametrów w podpróbkach

Model nie przeszedł testu Chowa, co wskazuje na brak stabilności parametrów w podpróbkach. Konsekwencje niespełnienia testu są takie same jak w przypadku testu RESET – nie jesteśmy w stanie udowodnić nieobciążoności i efektywności estymatora MNK, przez co wiarygodność interpretacji modelu jest podważona. W celu naprawienia tego problemu, powinniśmy stworzyć dwa oddzielne modele, które estymowałyby oddzielnie wyświetlenia filmów dodanych przed i po roku 2020. Alternatywą byłoby dodanie interakcji zmiennej *past2020* z każdą zmienną ciągłą w modelu.

W ostatnim teście zweryfikujemy czy składnik losowy ma rozkład normalny.

Nazwa testu	Testowane założenie KMRL	Hipoteza zerowa	Statystyka testowa	p-value	Decyzja weryfikacyjna
Test Jarque-Bera	Normalność rozkładu składnika losowego	Składnik losowy ma rozkład normalny	110.60	0	Odrzucamy H0 o normalności rozkładu składnika losowego na rzecz nienormalności rozkładu składnika losowego

Na podstawie testu odrzucamy hipotezę o normalności rozkładu składnika losowego. W przypadku modelu o liczbie obserwacji mniejszej niż 100 byłoby to problematyczne, ze względu na rozbieżność rozkładów statystyk w modelu od rozkładów standardowych. W modelu występują 364 obserwacje, przez co negatywna konsekwencja niespełnienia testu Jarque-Bera w nim nie występuje.

6. Analiza współliniowości i obserwacji nietypowych

W tym rozdziale na początku przyjrzymy się aspektowi współliniowości w modelu. W tym celu przeanalizujemy tabelę z współczynnikiem inflacji wariancji VIF, który określi nam, czy między badanymi predyktorami występuje współliniowość.

zmienna	VIF
Intercept	26.96
np.log(comments)	1.27
np.log(estimated_revenue)	1.06
vid_thumbnail_ctr_percent	1.25

Jedyna zmienną o wysokim współczynniku VIF jest stała, co jest normalnym zjawiskiem. W związku z tym problem współliniowości między zmiennymi objaśniającymi nie występuje w modelu. Jeśli jednak byłoby inaczej, należałoby sprawdzić istotność zmiennej o wysokim VIF>10, a następnie rozważyć jej usunięcie.

Obserwacjami, którym warto się przyjrzeć, to numery 0, 2 oraz 10. Każda z nich posiada wysoką wartość dźwigni, studentyzowanych reszt oraz odległość Cook'a.

Obserwacja	Długość filmu (s)	Data	Komentarze	CTR	Estymowany przychód ($)	Wyświetlenia
0	201	2016-06-02	91	27,66	0,561	23531
2	133	2016-06-14	0	7,07	0,089	6153
10	29	2016-08-17	0	5,13	0,231	6209

Każda z tych obserwacji dotyczy najstarszych filmów ze zbioru danych. Unikalność obserwacji 0 wynika z wartości CTR, która jest maksymalną zarejestrowaną wartością w zbiorze. W przypadku obserwacji 2 oraz 10, wspólną cechą jest brak komentarzy. Każda z tych obserwacji reprezentuje dość ekstremalne przypadki, jednak nie mamy podstaw do wyrzucenia owych

obserwacji. Nie są one spowodowane błędem danych, zatem poprawny model powinien być przygotowany na takie przypadki w przyszłości.

7. Podsumowanie

Model okazał się być niestety wysoce wadliwy. Nie udało się spełnić żadnego z kluczowych założeń klasycznego modelu regresji liniowej, przez co nie możemy interpretować jego oszacowań – możemy jedynie zakładać, że w pewnym stopniu odzwierciedla rzeczywistość. Mając to na uwadze, przejdźmy do weryfikacji hipotez. Główną hipoteza dotyczyła pozytywnego wpływu komentarzy na liczbę wyświetleń. Hipoteza została zweryfikowana pomyślnie - w każdej wersji modelu zmienna logarytmu liczby komentarzy była statystycznie istotna, z dodatnią wartością współczynnika. Hipotezy dotyczące większej popularności filmów o dłuższym czasie trwania oraz tych dodanych przed rokiem 2020 zostały odrzucone. Obie zmienne okazały się być nieistotne, a wręcz łącznie nieistotne z innymi zmiennymi nieistotnymi podczas zastosowania metody GEDS, co wskazuje na ich znikomy wpływ na wyświetlenia.

Jednym z głównych problemów modelu jest wysoka korelacja pomiędzy zmiennymi odpowiadającymi za zaangażowanie widowni. Były to czynniki, które sugerując się literaturą oraz własną intuicją zamierzałem wykorzystać, jednak nie byłoby to w naszym przypadku miarodajne. Alternatywą mogłaby być głębsza analiza tych filmów. Ponownie powołując się na literaturę, warto byłoby poszerzyć zbiór danych o informacje, które wymagałyby indywidualnej analizy każdego filmu. Czynniki takie jak typ filmu, lub produktywność autora mogłyby znacznie ubogacić model. Kolejnym sposobem na poprawę modelu jest analiza sentymentu każdego filmu przez sztuczną inteligencję – wskaźnik mierzący chociażby kontrowersyjność, nacechowanie emocjami bądź generalny sentyment filmu również mógłby przynieść poprawione rezultaty w przyszłości.

Bibliografia

Literatura:

Shen J., *The research of the factors that influence the popularity of YouTube videos,* Nexus International School, 2024

Velho R. M., Mendes A. M. F., Azevedo C. L. N., *Communicating Science With YouTube Videos: How Nine Factors Relate to and Affect Video Views, State University of Campinas, 2020*

Welbourne D. J., Grant W. J., *Science communication on YouTube: Factors that affect channel and video popularity,* 2015

Chatzopoulou G., Sheng C., Faloutsos M., *A First Step Towards Understanding Popularity in YouTube,* 2010

Zhou R., Khemmarat S., Gao L., Wan J., Zhang J., Yin Y., Yu J., *Boosting video popularity through keyword suggestion and recommendation systems*, 2016

Dane:

https://www.kaggle.com/datasets/positivealexey/youtube-channel-performance-analytics