Indeks	Punkty	Uwagi - sprawozdanie	Punkty (prezentacja)	Uwagi - prezentacja
258544	5	Jeśli mamy dane w złej skali, to łatwo wiele obserwacji uznać za outliery. Duży plus za znalezienie błędu w danych!	3	"Opis danych - ok Dokładny opis z czego korzystamy - ok. Język matematyczno-statystyczny: Szczegóły podział na 3 części - niepotrzebne. Przekształcenie danych (symetryzacja danych) Skośność, Kurtoza (na rysunku) Przycięcie zmiennych Uwaga: Można dać ładne etykietki na przekształconej skali Uproszczenie zmiennych - bardzo dobre. Ale czy należy to tłumaczyć osobom, które tego nie zrozumieją? Bardzo dużo szczegółów technicznego dochodzenia do rozwiązania. A nas interesuje: * jak wygląda model, który został wybrany * jak należy go rozumieć * na ile da się przewidywać liczbę udostępnień Bardzo dobry wykres 'roznica predykcji i przeczywistych' vs 'indeksy w sortowaniu' Jak poprawić -> dobre pomysły.
258486, 258	5	Przy wypisywaniu tabel warto użyć pakietu xtable lub knitr::kable	5	"Cel projektu -> w miarę jasno opisane Opis danych: krótko jakie są grupy zmiennych Eksploracja: identyfikacja potencjalnych problemów z danymi.
258478	4	Dobre uzasadnienie czemu bierzemy skalę logarytmiczną Co do wyboru ostatecznego modelu: co prawda R^2 jest niskie w obu przypadkach 0.08 i 0.12, ale to jest jednak różnica prawie 50%! Brak diagnostyki modelu	4	"Początek dobry: nie ma nic super, ale mamy prosty model, który trochę tłumaczy. Bardzo dobre wytłumaczenie co się dzieje w modelu. Brak opisu danych. Opis - kilka hitów, ale większość średniaków. Przejście do innej skali (logarytm to zgrzyt, ale malutki) Mini minus - warningi w chunkach warning=FALSE, message=FALSE Jeśli mamy średnie dla dwóch grup - można sprawdzić czy warto w ogóle raportować

258554	4	Podział zbioru na treningowy, walidacyjny i testowy powinien nastąpić PRZED eksploracją danych! Usunięcie obserwacji z bardzo dużą liczbą udostępnień jest ok. Usunięcie tych z małą liczbą może być niebezpieczne. Warto sprawdzić jak model zachowuje się dla takich obserwacji Usunięcie obserwacji 29088 ok Jeśli chodzi o usunięcie obserawcji z dużą liczbą linków - jest to nieco pochopne. Dlaczego one nam przeszkadzają? Może lepiej zastosować transformację? Połączenie poziomów zmiennej channel ok	5	"Początek: Ok. Co robimy, co możemy dostać. Opis danych - bardzo dobrze. Analiza udostępnień - co za tym stoi - ciekawe. Dużo niepotrzebnych zmiennych - będziemy się chcieli ich pozbyć. Czyszczenie danych: obserwacje zbyt zróżnicowane (może mogłaby pomóc transformacja?), usuwane są obserwacje odstające (uwaga, czy to nie są przypadkiem te najbardziej popularne?) Troszkę przesada z łączeniem grup. Wchodzimy w szczegóły - w jakim celu.
258512, 258	4.5	Podział zbioru na treningowy, walidacyjny i testowy powinien nastąpić PRZED eksploracją danych! Bardzo fajny wykres shares vs all Usunięcie obserwacji ze względu na błędne wartości n_tokens_title bardzo ok Właściwie zamiast korelacji między zmiennymi objaśniajacymi powinniśmy mierzyć Variance Inflation Factor (VIF), który mówi w jakim stopniu zmienna jest tłumaczona liniow przez wszystkie pozostałe Co do modelu model1_lm_log - widać niespełnienie założeń. I widać też, że powodem tak naprawdę jest kilka obserwacji o bardzo dużej liczbie udostępnień Bardzo fajny pomysł z porównaniem artykułów parami	5	"Początek ok. Cel, co robimy i po co. Opis danych - z czym mamy do czynienia, grupy zmiennych. Jak wyglada zmienna objasniana - mamy bardzo skrajne różnice w udostępnieniach. Większość zmiennych nie ma związku - pokazujemy to co ma wpływ na liczbę udostępnień. Tytuły -> średnia długość tytułu daje szanse na dużo sharów Fajna idea odnośnie porównań między dwoma artykułami. Ale przy posortowaniu mamy problem wielokrotnego testowania. Wnioski też ok -> zamiast przewidywać liczbę sharów czy jest popularny czy nie.