Analiza odchodzenia klientów

Bartosz Chądzyński & Michał Turek

2023-05-03

Wstęp

Preprocessing

Wykresy

Zacznijmy od analizy wykresów. Na początek zmienne ciągłe. Na wykresie 1 i 3 widzimy, że zmienne te są w znacząco różnych skalach, więc prawdopodobnie potrzebna będzie normalizacja. Natomiast na wykresie 2 i 4 widać, że każda ze zmiennych ma istotnie różny rozkład, gdy pogrupujemy ją ze względu na Churn.

Z kolei na 5 widzimy, że w niektórych przypadkach są duże różnice w ilości obserwacji z każdej kategorii, jeśli chodzi o daną zmienną. W szczególności takimi zmiennymi są *PhoneService*, czy *MultipleLines*.

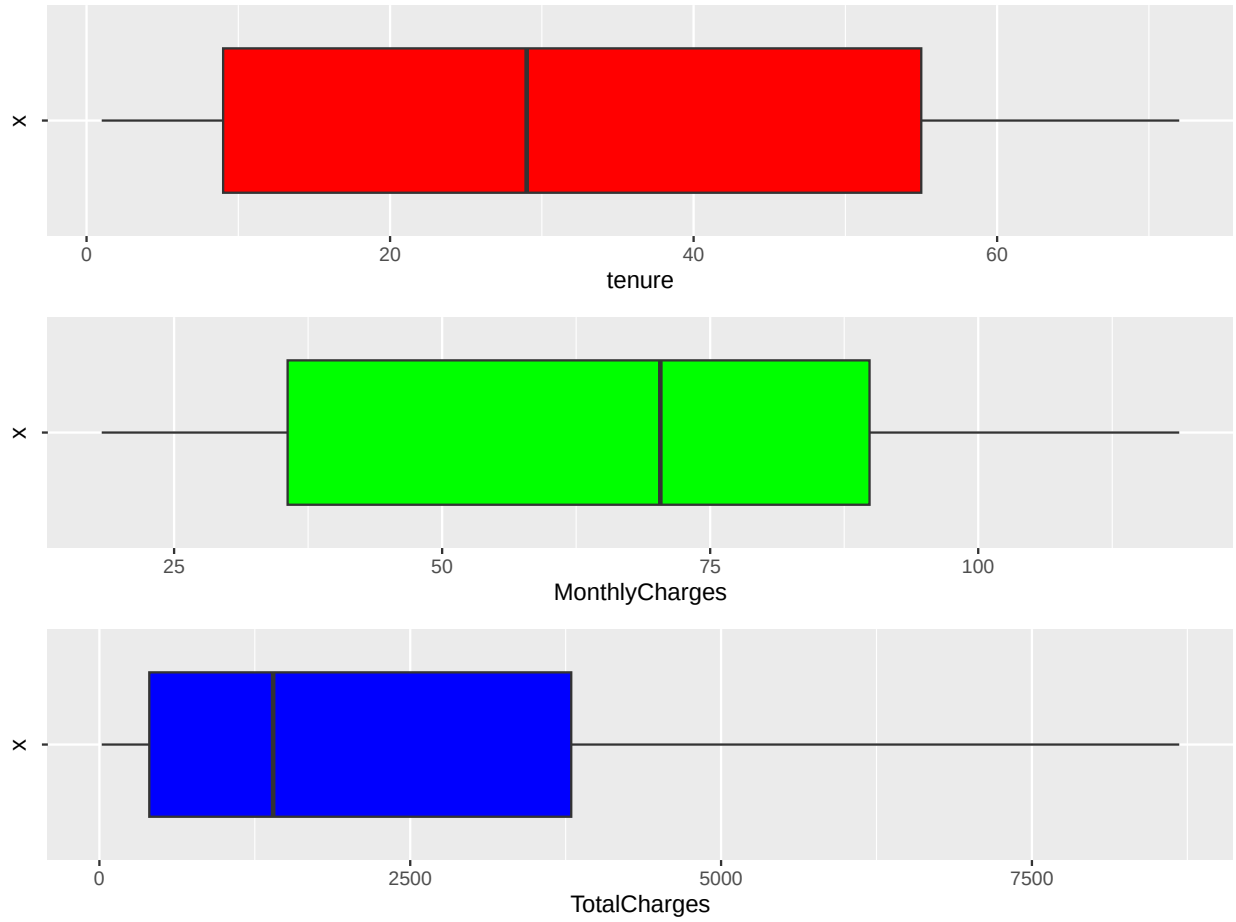

Figure 1: Boxploty zmiennych ciągłych

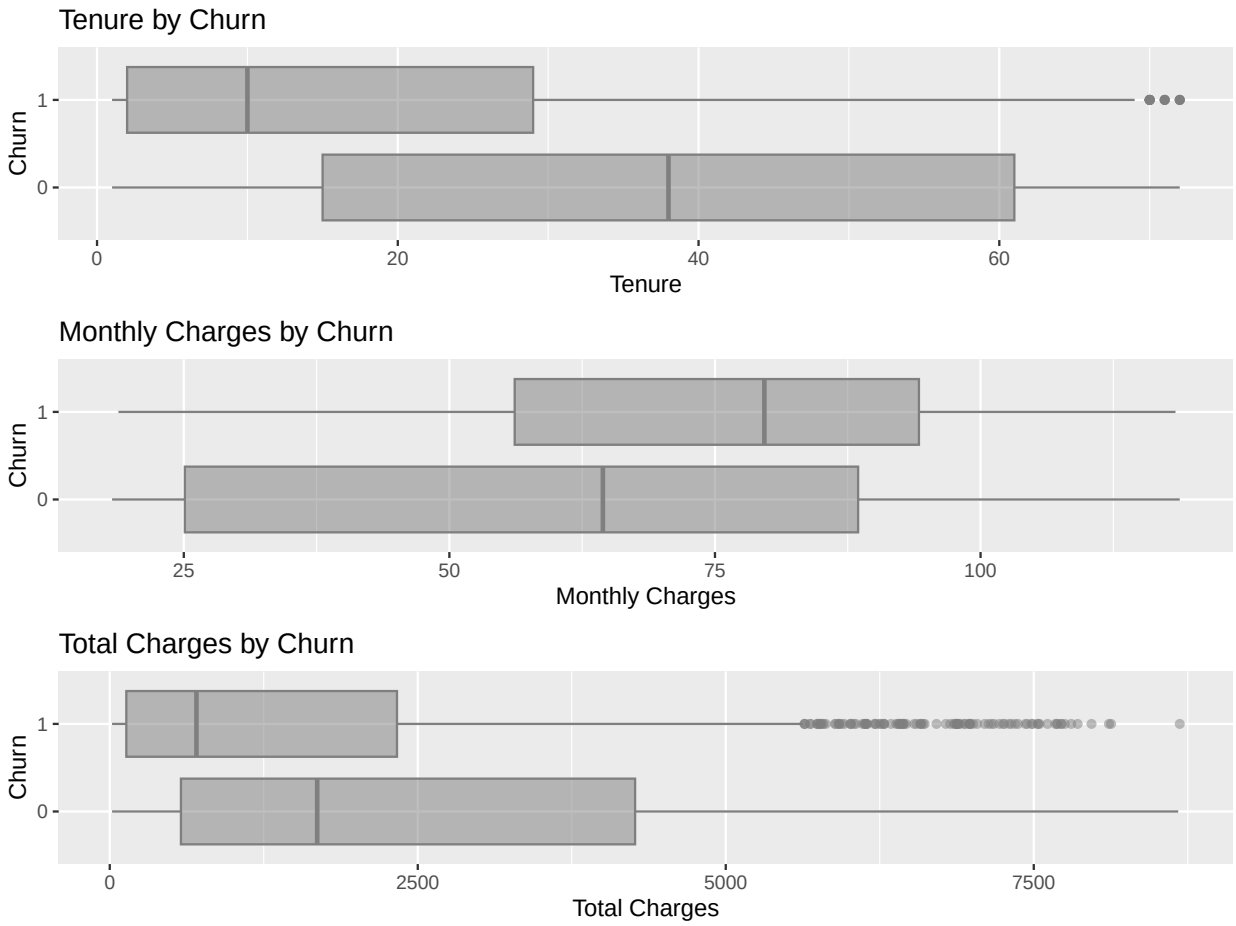

Figure 2: Boxploty zmiennych ciągłych z podziałem ze względu na Churn

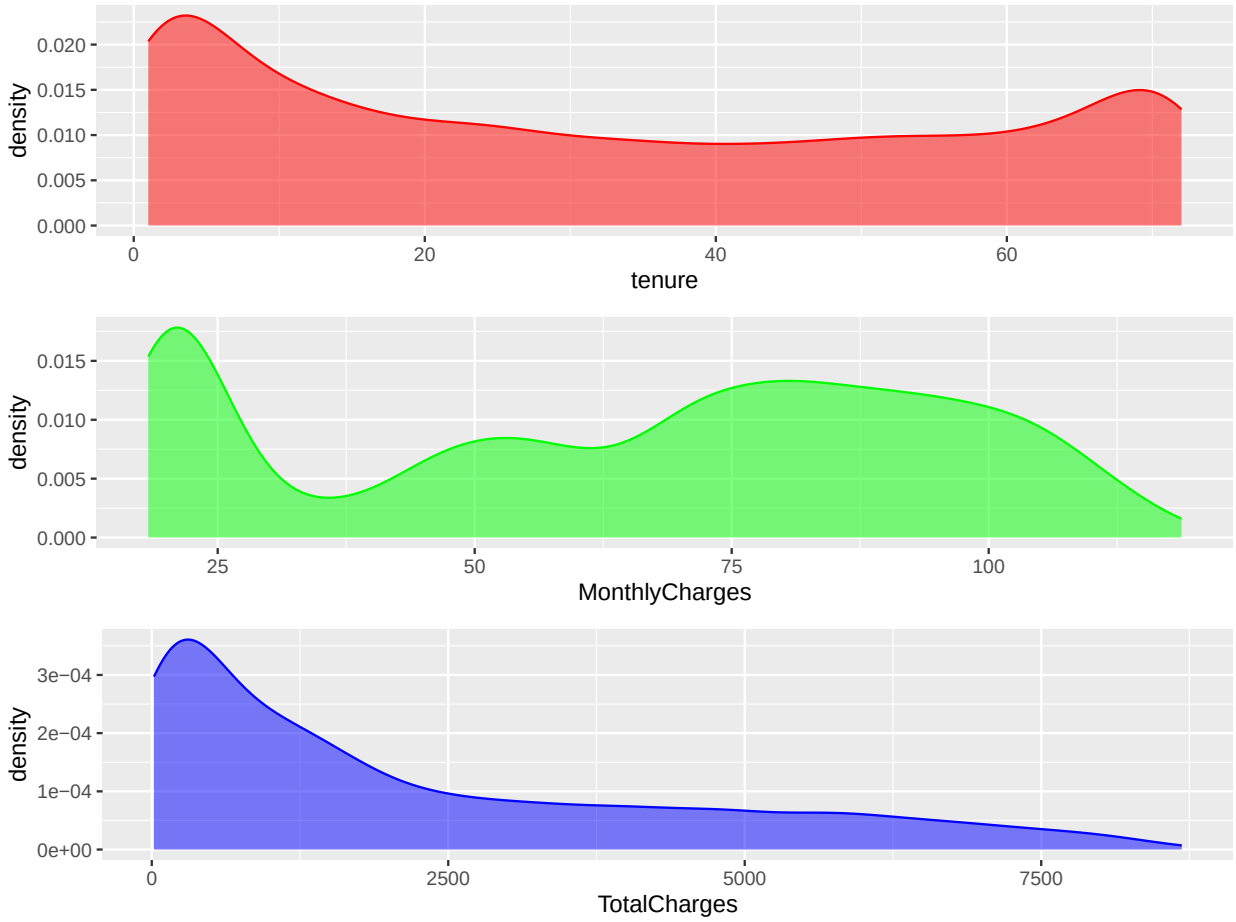

Figure 3: Estymator jądrowy gęstości

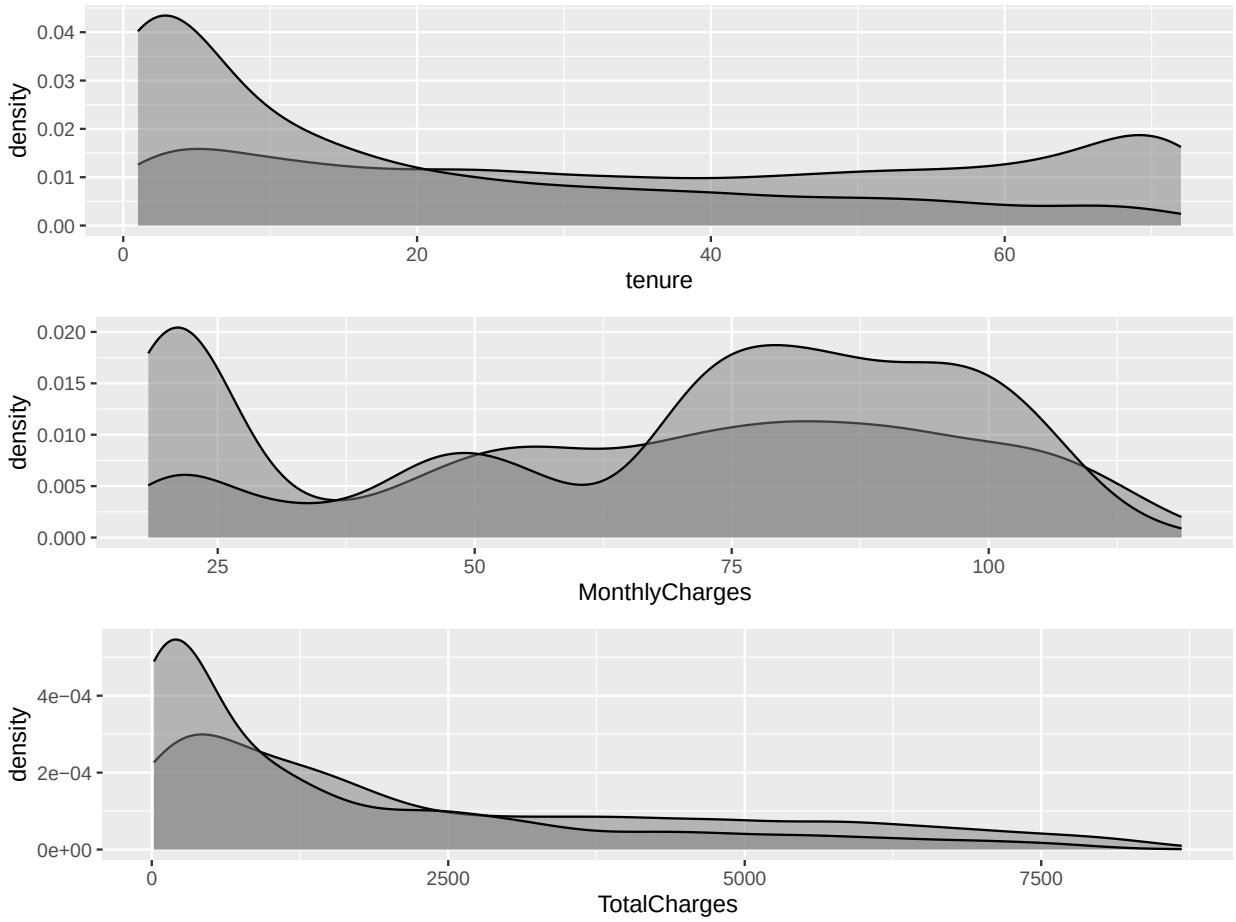

Figure 4: Estymator jądrowy gęstości z podziałem ze względu na Churn

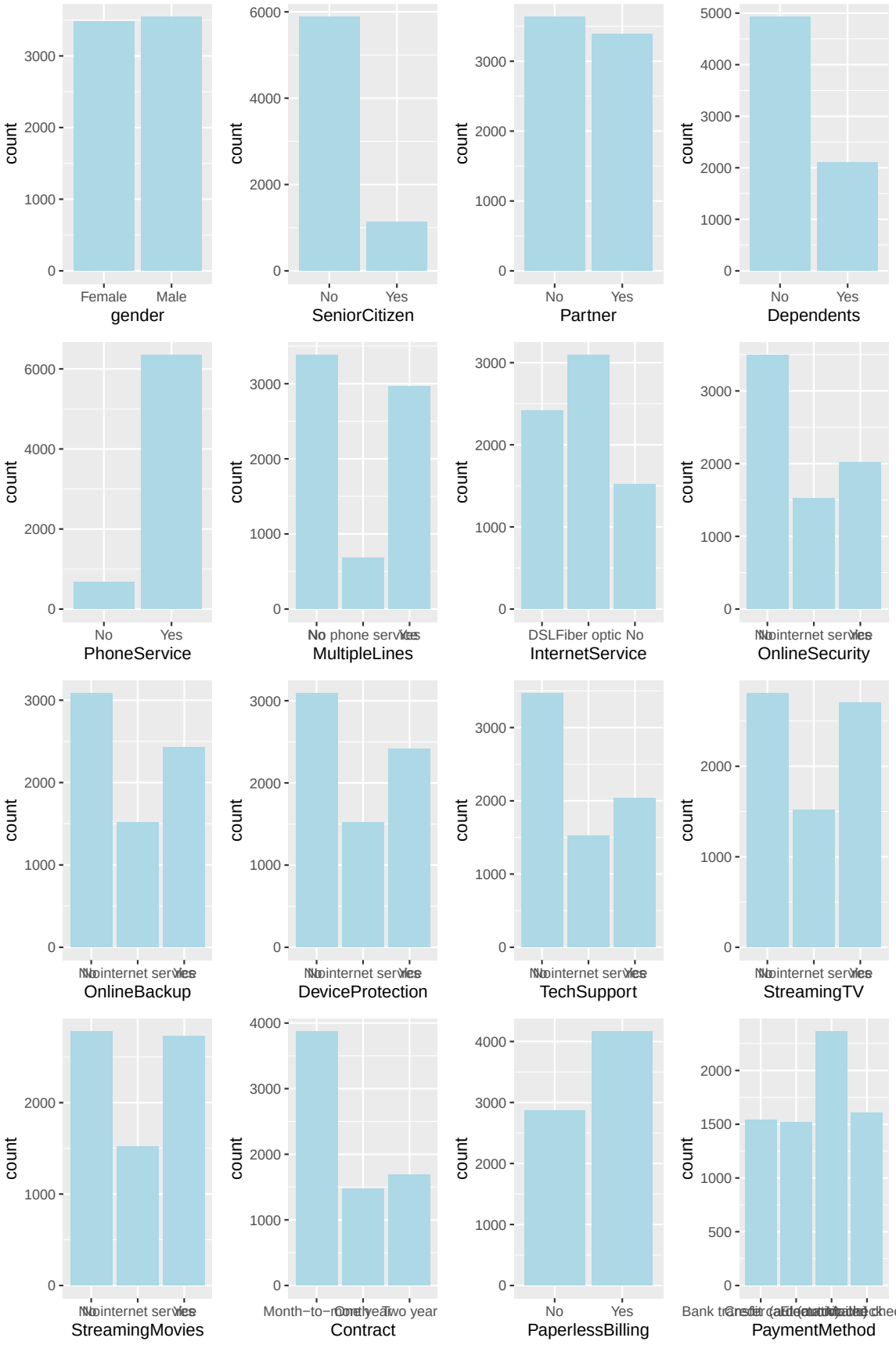

Figure 5: Wykres ilości obserwacji z podziałem na kategorie zmiennych

Interpretacja Wyników

W naszych danych jest zaledwie 11 obserwacji z brakującymi danymi (na 7033 łącznie). Zatem zasadne jest pominięcie ich w trakcie analizy danych. Nie stosujemy żadnej imputacji. Ilość danych może być obciążająca dla niektórych modeli. Jeśli będą występowały problemy ze złożonością obliczeniową, to dla konkretnego modelu będziemy decydować o przeprowadzeniu analizy dla ewentualnego podzbioru danych.

W tabeli poniżej mamy macierz korelacji zmiennych ciągłych. Jak widać istnieje mocna korelacja pomiędzy tym jak długo klient korzysta/korzystał z usług, a kwotą jaką zapłacił za usługi. Nie powinno to dziwić. Na razie jednak nie decydujemy się na wyrzucenie którejś ze zmiennych, ponieważ zarówno czas jak i koszt może być istotny w kontekście odchodzenia klientów. Te dwie rzeczy nie muszą być ze sobą powiązane w pełni. Może być tak, że odchodzą głównie nowi klienci, niezależnie od tego ile płacą. Albo może być tak, że odchodzą klienci, którzy zapłacili rachunki powyżej pewnej sumy, niekoniecznie będący długo/krótko stażem.

	tenure	MonthlyCharges	TotalCharges
tenure	1.00	0.25	0.83
MonthlyCharges	0.25	1.00	0.65
TotalCharges	0.83	0.65	1.00

Najprawdopodbniej potrzebne będzie wykonanie transformacji danych, w szczególności normalizacji. Natomiast jeśli chodzi o obserwacje odstające, to nie ma ich za dużo. Natomiast pojawiają się licznie w przypadku zmiennej *TotalCharges* pogrupowanej ze względu na *Churn*. Widać, że jest tendencja, aby odchodzący klienci należej do jednej z dwóch grup. Są albo nowymi klientami, albo klientami z dużym stażem. Ta druga grupa jest na wykresie pudełkowym interpertowana jako obserwacje odstające. W rzeczywistości należy to interpretować tak, że rozkład tej zmiennej jest dwumodalny,nie będziemy stosować technik mających na celu ignorowanie lub zmniejszenie wpływu tych obserwacji, znacząco odbiegających od reszty.