Raport 2 ZML

Erwin Jasic

14 kwietnia 2021

Cel raportu:

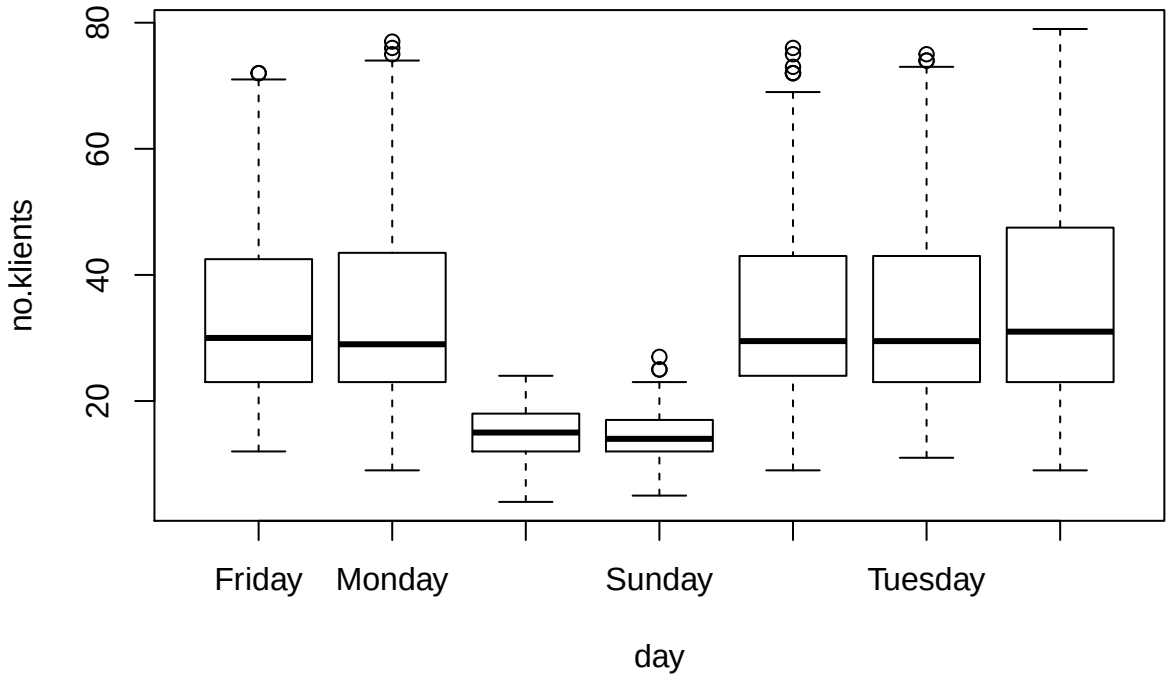
W raporcie zajmiemy się zastosowaniem teorii z wykładu dotyczącego regresji poissona na zbiorze danych sklep. Zbiór ten opisuje liczbę klientów w poszczególnych dniach tygodnia, godzinach oraz zawiera informacje o tym czy w tym dniu miało miejsce wydarzenie sportowe.

Zadanie 1

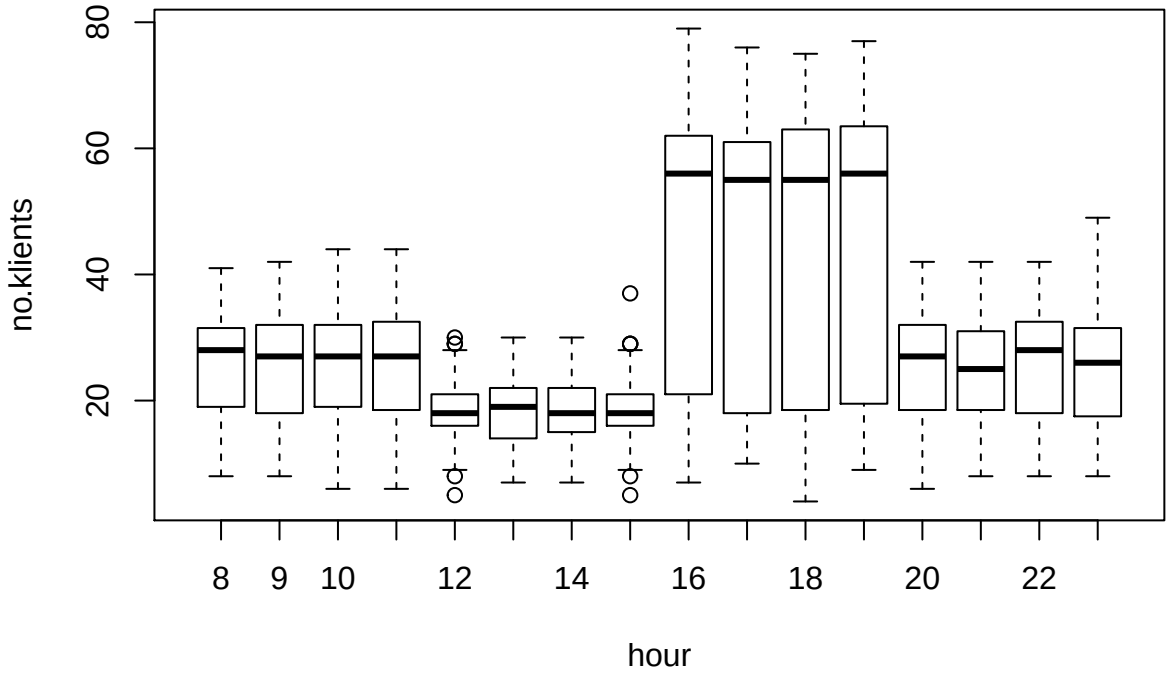
Wczytujemy dane.

Zadanie 2

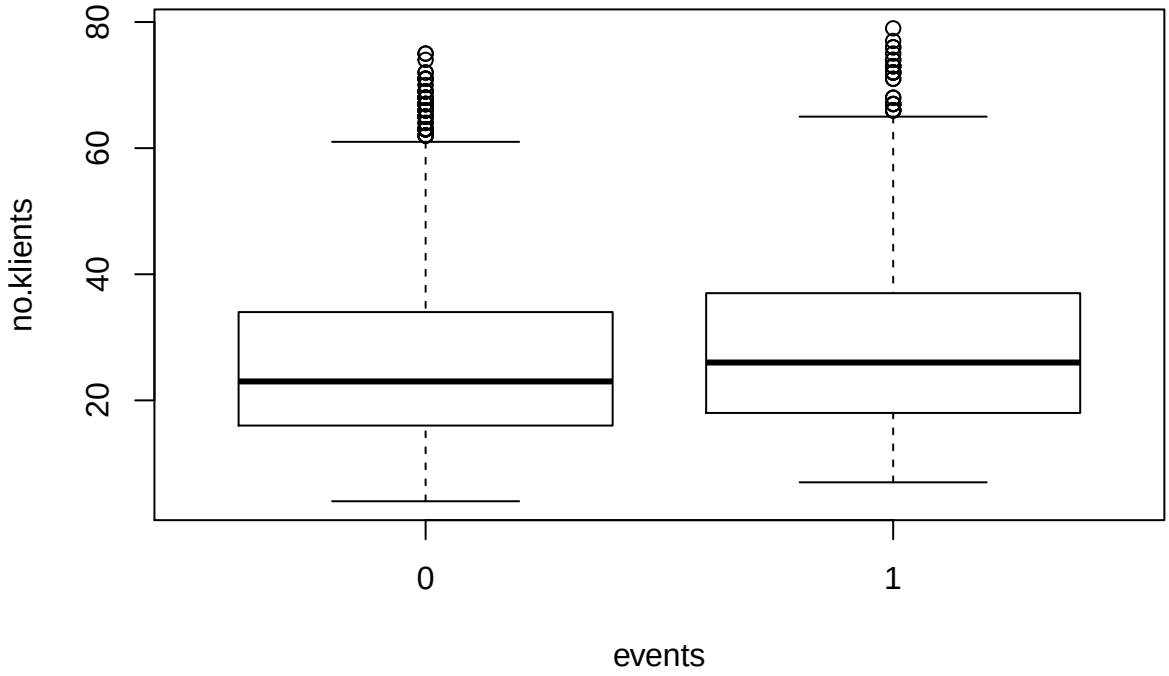
W tym zadaniu mamy dokonać wstępnej analizy tego zbioru danych poprzez tworzenie wykresów. Puścimy po kolei funkcje z tresci zadania i postaramy się na podstawie uzyskanych wykresów wyciągnąć wnioski.



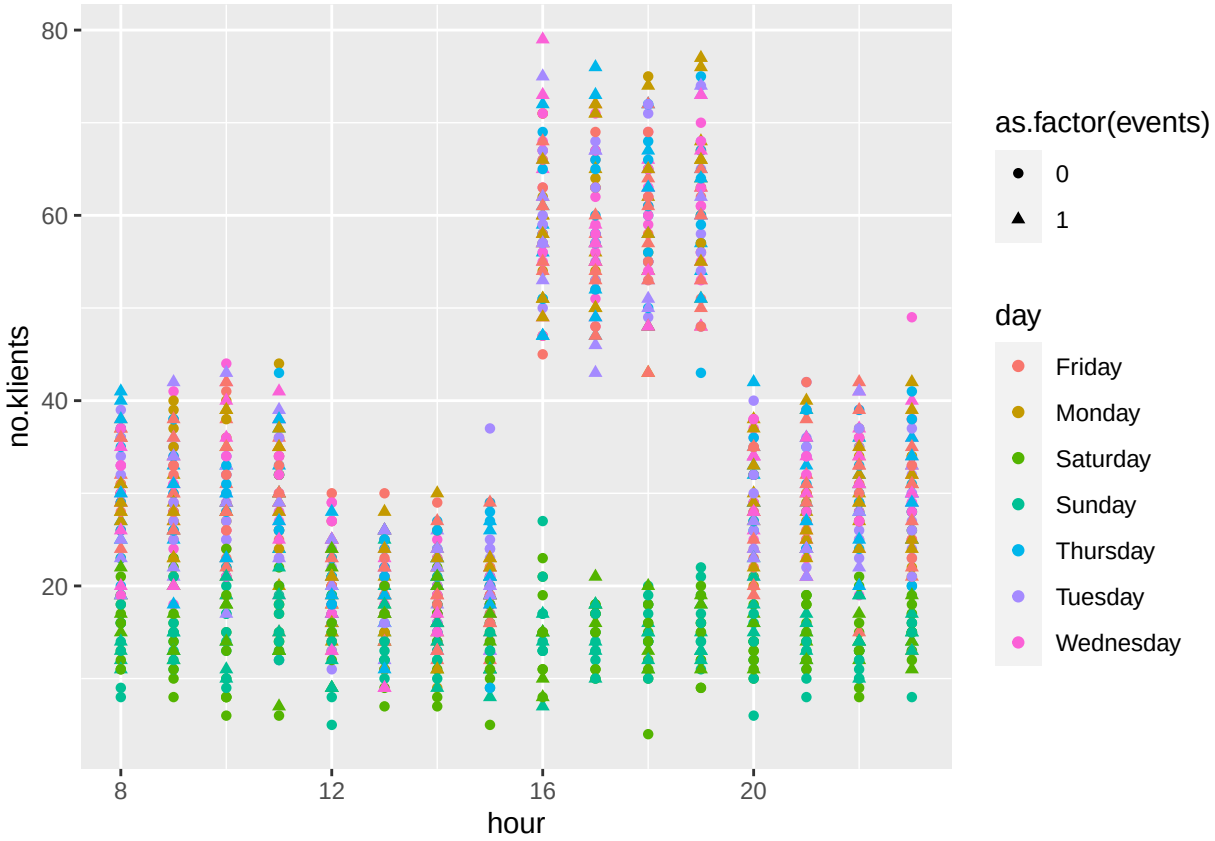
Na podstawie tego boxplota łatwo możemy zauważyć, że na weekendy do sklepu przychodzi dużo mniej osób, a w dni robocze mniej więcej tyle samo niezależnie od tego, który jest to dzień roboczy..



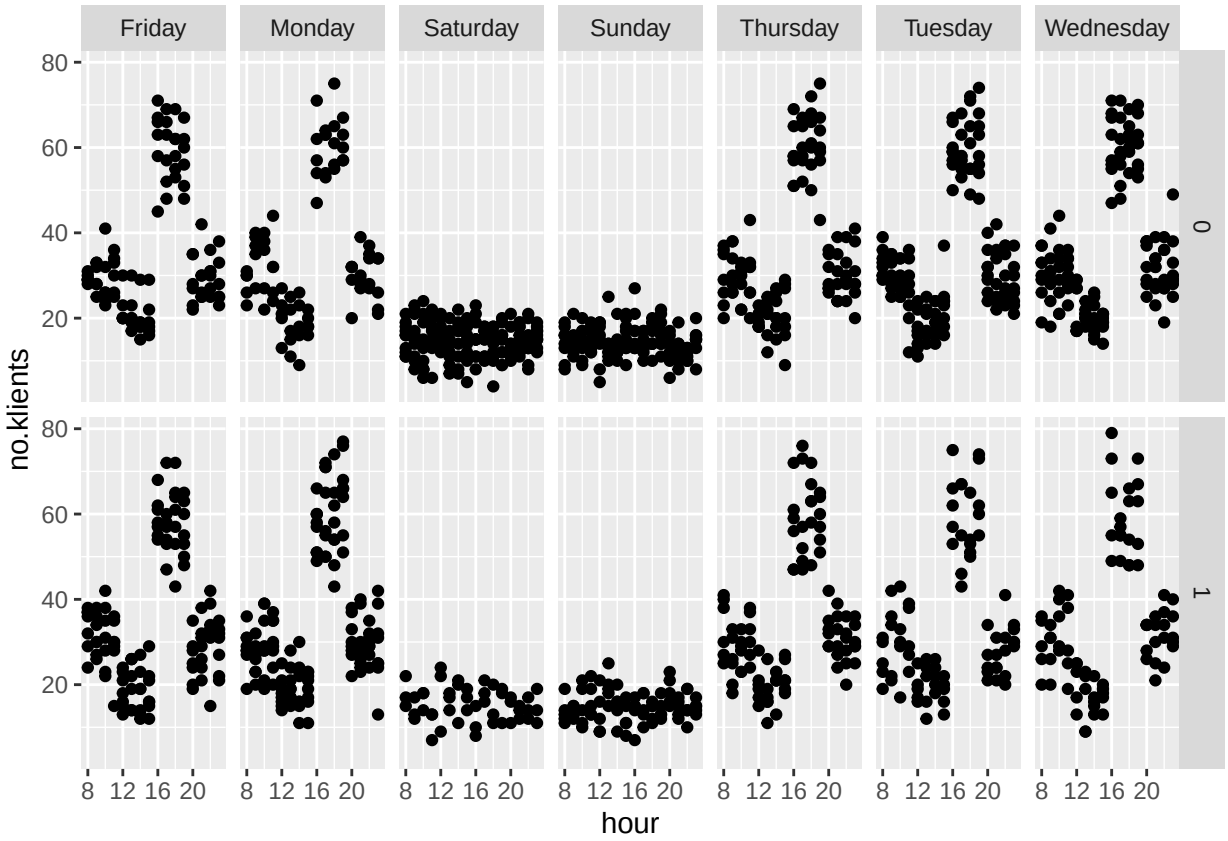
Tutaj możemy zaobserwować, że najmniej osób przychodzi między godziną 12 a 15, a najwięcej miedzy 16, a 19.



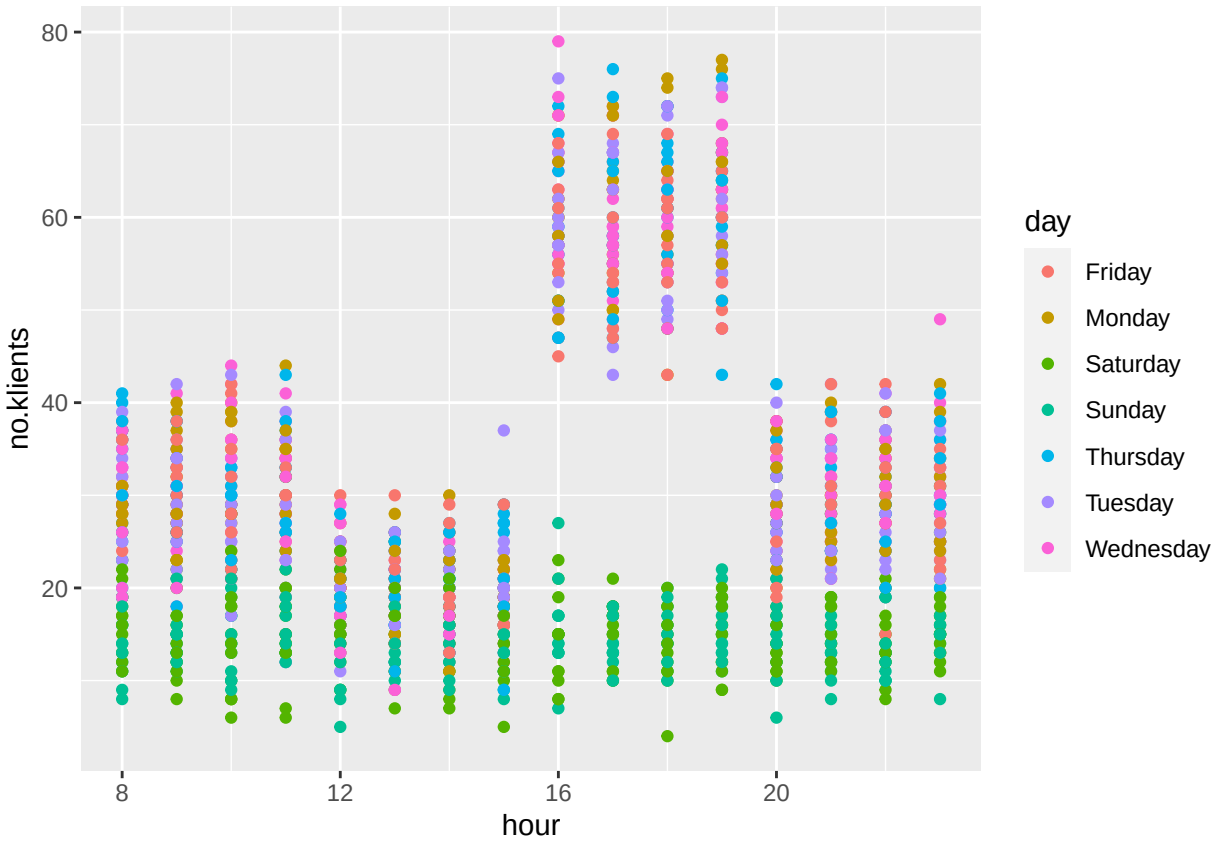
Ten wykres pokazuje zależność między liczbą klientów, a tym czy w danym dniu ma miejsce jakies wydarzenie sportowe. Jak widzimy na powyższym boxplocie, nie ma żadnej zależności.



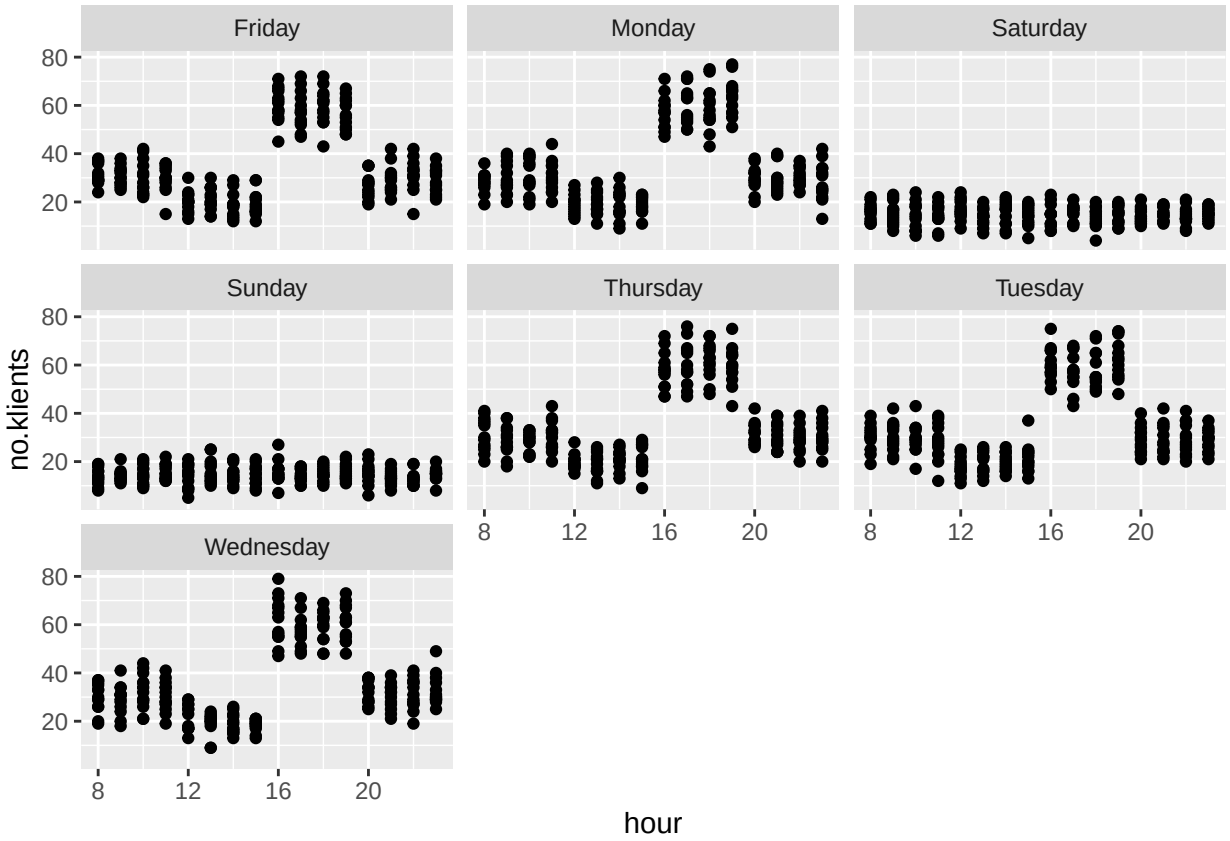
Z tego wykresu możemy wysnuć podobne wnioski jak ze wszystkich wykresów powyżej. Ciekawą rzecz jaką możemy zaobserować to, że od 16 do 19 godziny są największe różnice między dniem roboczym a weekendowym.



Na podstawie tych rozrzutów widzimy, że parametr events nie ma wpływu na to ilu klientów przychodzi danego dnia (rozkłady są podobne w tych samych dniach).



Tutaj mamy ten sam wykres co dwa wyżej, ale bez uwzględnienia parametru events. Cieżko odczytać z tego wykresu coś więcej niż do tej pory.



Ten wykres mówi nam sporo o rozkładzie ludzi w ciągu dnia w poszczególnych dniach tygodnia. Widzimy ze od 16 do 19 najwięcej ludzi w ciągu tygodnia (poniedziałek - piatek), na weekend mniej więcej tyle samo ludzi na kazda godzinę.

Zadanie 3

W tym zadaniu skonstruujemy model Poissona z interakcją pomiędzy wszystkimi regresorami i potraktujemy te regresory jako faktory. Po tej konstrukcji przeprowadzimy krótką analizę i odpowiemy na pytania zawarte w treści zadania.

Na podstawie p-wartości dla zmiennej/zmiennych events w modelu z interakcja jak i bez interakcji widzimy, że nie jest ona istotna na poziomie istotności 95%. Zgadza nam się to ze wstępną analizą z zadania 2, gdzie nie zauważyliśmy róznic na wykresach między dniamy, gdzie było jakieś wydarzenie sportowe oraz gdy takiego nie było. Zmiennych w tym modelu mamy 25 (16 godzin + 7 dni + 2 events) nie licząc liczby klientów. Wszystkich wierszy mamy 1456, a kolumn 4.

Zadanie 4

W tym zadaniu chcemy zmniejszyć liczbę zmiennych. W tym celu grupujemy godziny na 4-godzinne bloki oraz oznaczamy dni tygodnia jako roboczy lub jako weekendowy. Tak stworzony model porównamy z tym z zadania 3.

```
## Analysis of Deviance Table
##
## Model 1: no.klients ~ hour + day + events + hour:day + hour:events + day:events
## Model 2: no.klients ~ weekend + pora_dnia + weekend:pora_dnia
##   Resid. Df Resid. Dev   Df Deviance Pr(>Chi)
```

```
## 1      1322     1457.7
## 2      1448     1552.5 -126  -94.735   0.9829
```

Nasz nowy model ma 6 (2 weekend + 4 pora_dnia) zmiennych. Wykonaliśmy test na to czy nasz nowy model różni się od pełnego modelu z zadania 3. Posługując się funkcją 'anova()' widzimy, że zredukowany model nie różni się od pełnego. Zatem udało nam się zredukować model w taki sposób, że nie straciliśmy istotnych informacji na jego temat.

Zadanie 5

	roboczy 8:00-11:59	roboczy 12:00-15:59	roboczy 16:00-19:59	roboczy 20:00-23:59
Średnia klientow na godzine	30	19.7	59.6	30
Postać predyktora	b0 + b4	b0 + b1 + b4 + b5	b0 + b2 + b4 + b6	b0 + b3 + b4 + b7
Wartość predyktora	3.4	2.98	4.09	3.4

	weekend 8:00-11:59	weekend 12:00-15:59	weekend 16:00-19:59	weekend 20:00-23:59
Średnia klientow na godzine	14.8	15	14.9	14.4
Postać predyktora	b0	b0 + b1	b0 + b2	b0 + b3
Wartość predyktora	2.69	2.7	2.7	2.67

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{1i} X_{4i} + \hat{\beta}_6 X_{2i} X_{4i} + \hat{\beta}_7 X_{3i} X_{4i}$$

, gdzie zmienne przyjmują wartości:
$X_{1i} = 1$, gdy i-ta obserwacja była zaobserwowana w godzinach 12:00-15:59, 0 wpp
$X_{2i} = 1$, gdy i-ta obserwacja była zaobserwowana w godzinach 16:00-19:59, 0 wpp
$X_{3i} = 1$, gdy i-ta obserwacja była zaobserwowana w godzinach 20:00-23:59, 0 wpp
$X_{4i} = 1$, gdy i-ta obserwacja była zaobserwowana w dniu roboczym, 0 wpp

Na podstawie tej tabeli widzimy, że nasze wstępne obserwacje były trafne, tzn. klienci odwiedzają sklep mniej więcej z tą samą częstotliwościa w ciągu dnia podczas weekednu, a z kolei w dzień roboczy największy ruch jest od godziny 16:00 do 19:59, a najmniejszy od 12:00 do 15:59.

Zadanie 7

Ostatnią rzeczą jaką musimy zrobić jest stworzenie grafiku pracy na podstawie tabeli z zadania 5. Zakładamy, że jeden pracownik jest w stanie obsłużyć do 20 klientów na godzinę.

	Dzień roboczy	Dzień weekendowy
8:00-11:59	2	1
12:00-15:59	1	1
16:00-19:59	3	1
20:00-23:59	2	1

Podsumowanie:

Dzięki wykonaniu poleceń dotyczących regresji Poissona widzimy, że mamy wiele róznych technik do porównywania ze sobą modeli. Widzimy, że nie zawsze model pełny jest tym optymalnym jaki powinniśmy analizować, ponieważ po zredukowaniu/pogrupowaniu zmniennych nie stracilismy za wiele informacji o rozkładzie danych, ale zyskaliśmy na przejrzystości. Dzięki takim zabiegom nawet duże zbiory danych, które mają wiele zmiennych można po uprzednim pogrupowaniu dość łatwo analizować i wyciągnać istotne wnioski.