Zaawansowane modele liniowe raport nr 2

Dominik Mika

19 kwietnia 2021

Zadanie 1

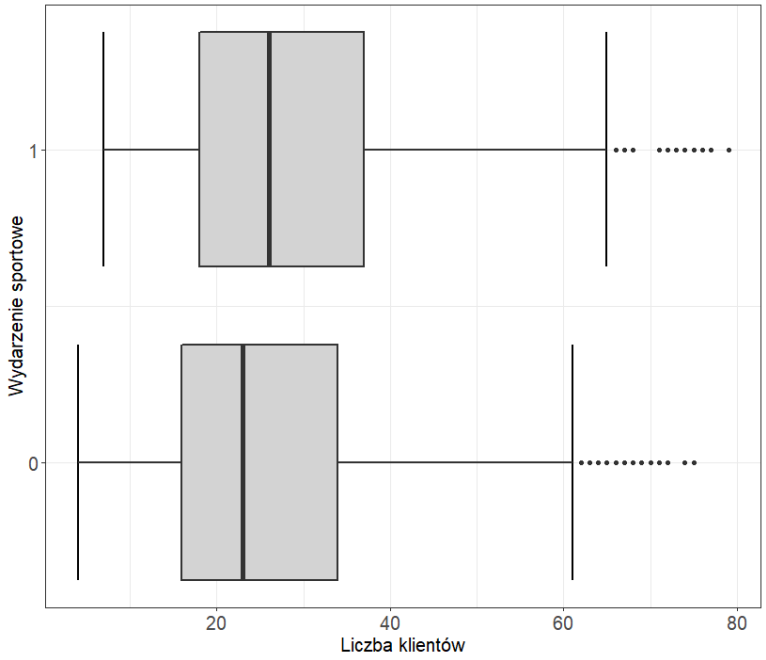
Na tej liście będziemy pracować ze zbiorem danych sklepu z pliku **sklep.txt**. Znajdują się w nim informacje o liczbie klientów przychodzących do pewnego sklepu w okresie około 3–ech miesięcy. W zbiorze znajdują się cztery kolumny:

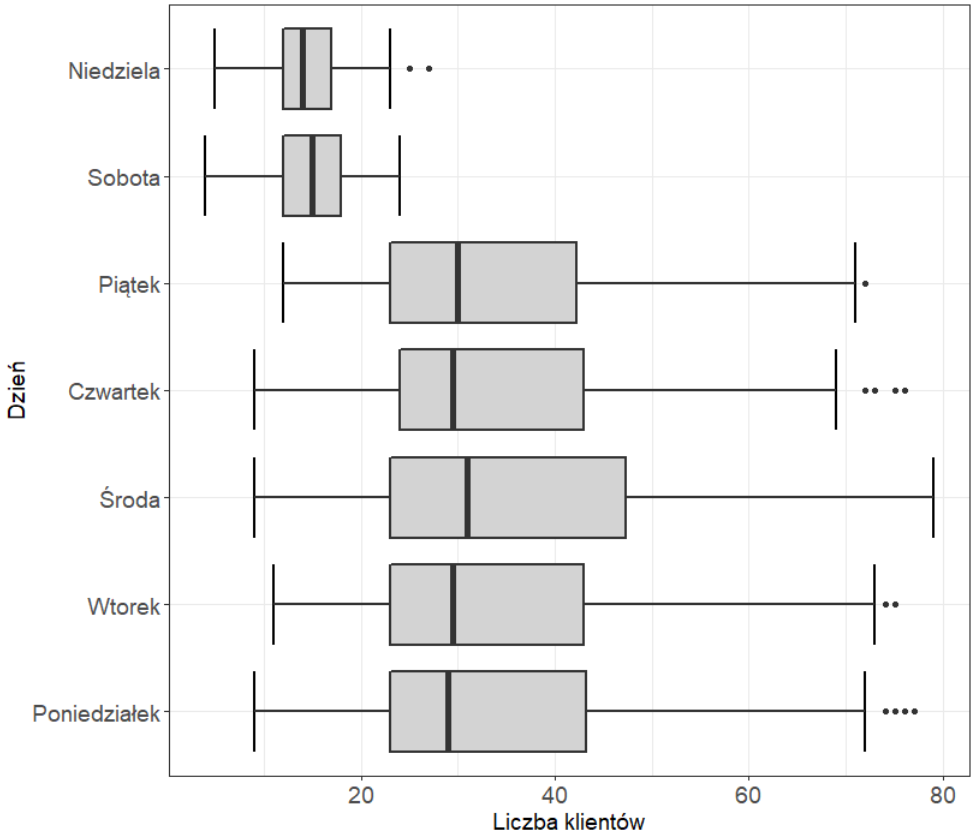
- no.klients – liczba klientów obsłużonych w danej godzinie (wart.: 0,1,2,...),
- day – dzień tygodnia (wart.: poniedziałek, wtorek,..., niedziela),
- hour – godzina (wart.: 8,9,...,23),
- events – informacja o tym czy w danym dniu miało miejsce jakieś wydarzenie sportowe (wart.: 0 - nie, 1 - tak)

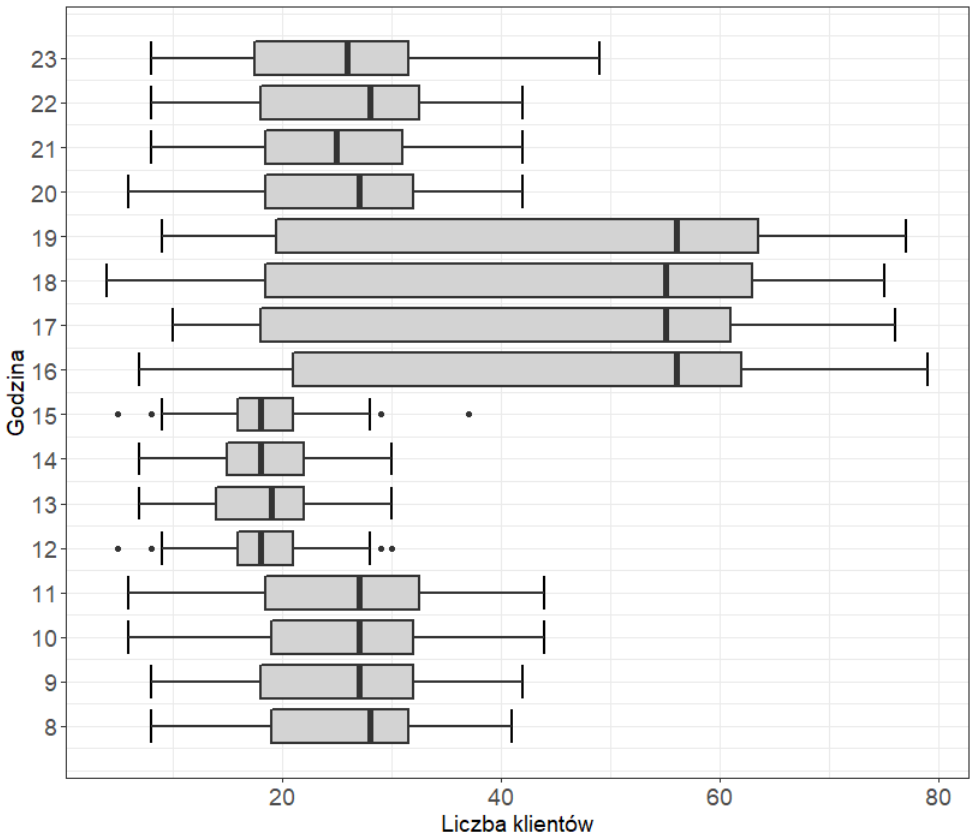
W poniższych zadaniach dokonamy analizy tego zbioru przy użyciu regresji Poissona, gdzie zmienną objaśnianą będzie liczba obsłużonych klientów, a pozostałe zmienne będą regresorami.

Zadanie 2

W tym zadaniu dokonamy wstępnej analizy tego zbioru poprzez wizualizację. Poniżej znajdują się wykresy typu boxplot dla poszczególnych zmiennych.


Wydarzenie sportowe
1
0
20
40
60
80
Liczba klientów


Dzień
Niedziela
Sobota
Piątek
Czwartek
Środa
Wtorek
Poniedziałek
20
40
60
80
Liczba klientów



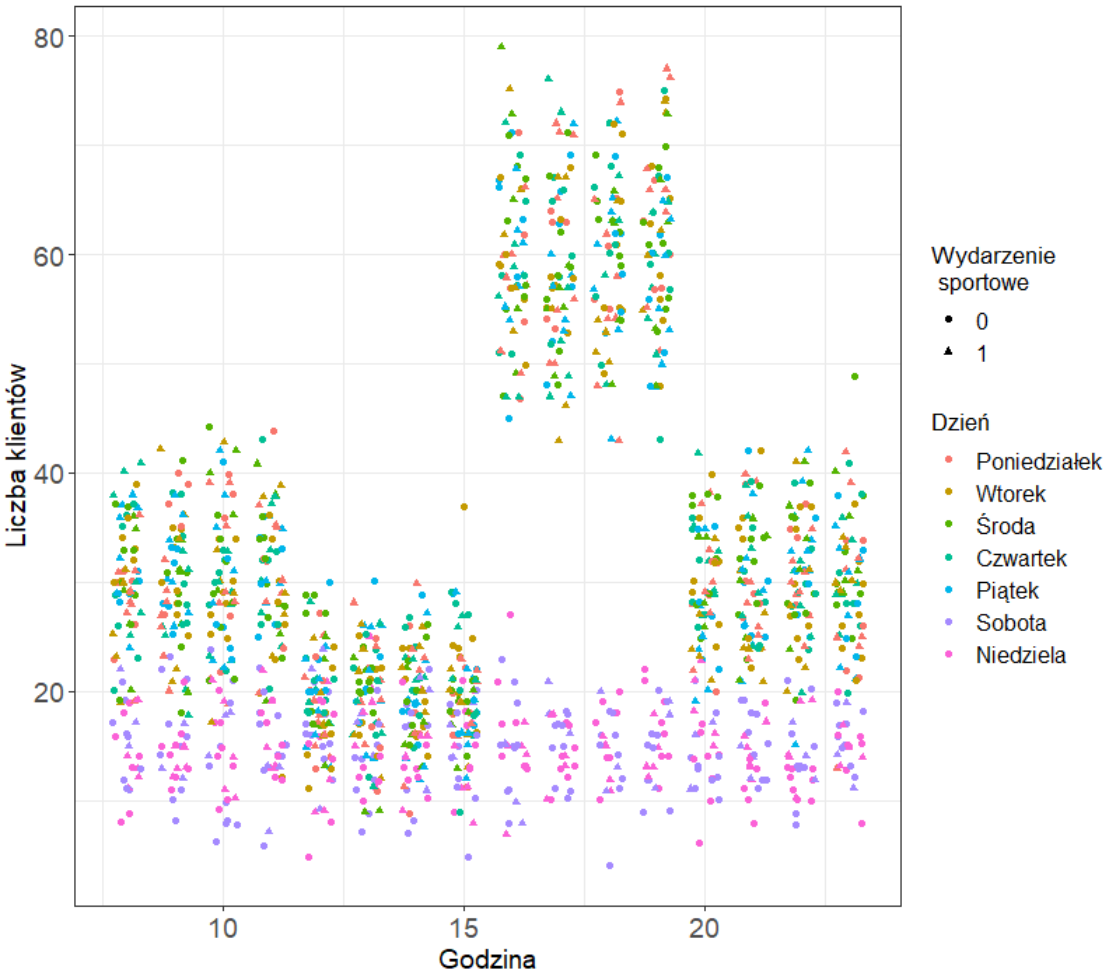
Po spojrzeniu na wykresy pudełkowe możemy zauważyć pewne trendy i własności. Po pierwsze widzimy, że wykresy pudełkowe w rozbiciu na zmienną *wydarzenie sportowe* nie różnią się od siebie znacząco. Możemy podejrzewać, więc że nie ma ona wpływu na liczbę klientów.

Dodatkowo na wykresie pudełkowym w rozbiciu na różne wartośc zmiennej *dzień* możemy zauważyć pewną zależność. Mianowicie w dni robocze liczba klientów wydaje się być podobna, w dni weekendowe podobnie.

Natomiast na wykresie dla zmiennej *godzina* obserwujemy, że liczba obsłużonych klientów układa się w 4-o godzinne bloki.

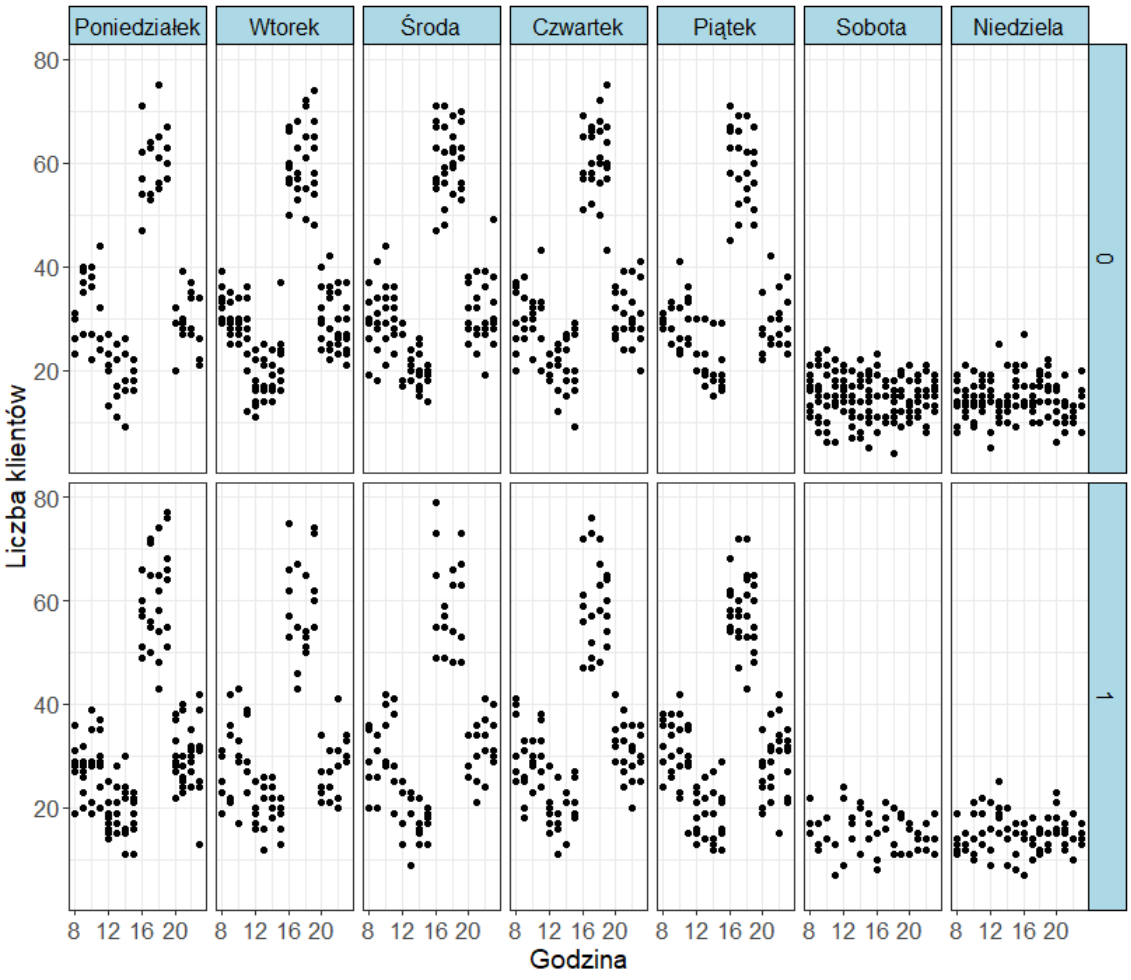
Spróbujmy spojrzeć jeszcze na inne wykresy.

Poniżej znajduje się wykres rozrzutu danych w zależności od dnia tygodnia oraz wydarzenia sportowego.



Widzimy, że zmiana kształtu punktów wydaje się być losowa. Po raz drugi widzimy, że zmienna *wydarzenie sportowe* prawdopodobnie nie wpływa na liczbę klientów.

Na poniższym wykresie znajdują się rozrzuty liczby klientów w zależności od godziny, w rozbiciu na różne dni tygodnia i wydarzenie sportowe.



Na powyższym wykresie obserwujemy takie same zachowanie zmiennych jakie podejrzewaliśmy wcześniej.

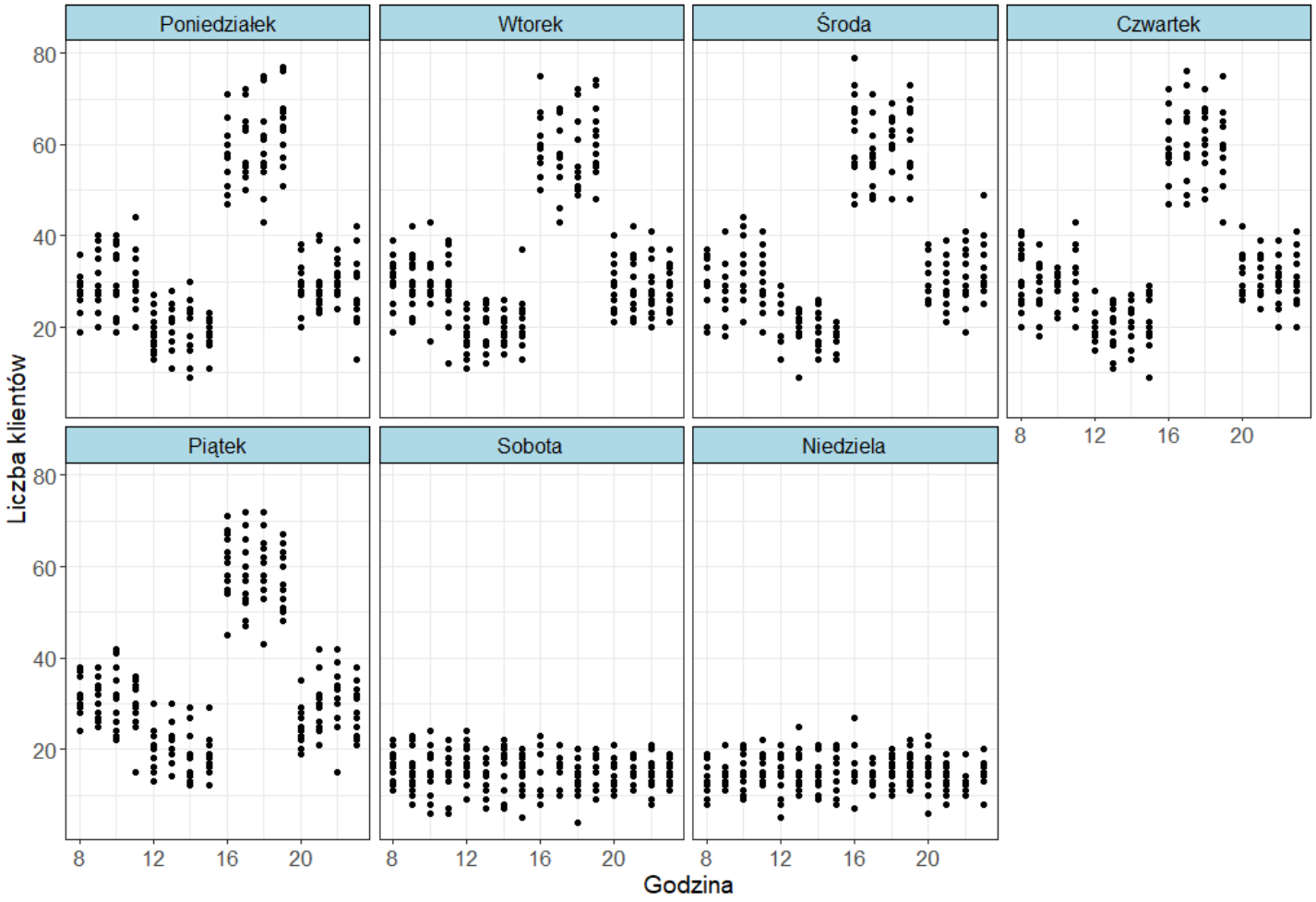
Po wstępnej obserwacji danych możemy zauważyć, że zbiór po rozbiciu ze względu na *wydarzenie sportowe* nie różni się. Ewidentnie mamy mniej obserwacji, gdy ta zmienna przyjmuje wartość 1, ale nie widzimy żadnych zmian w liczbie osłużonych klientów. Wszystkie powyższe wykresy nam to sugerują.

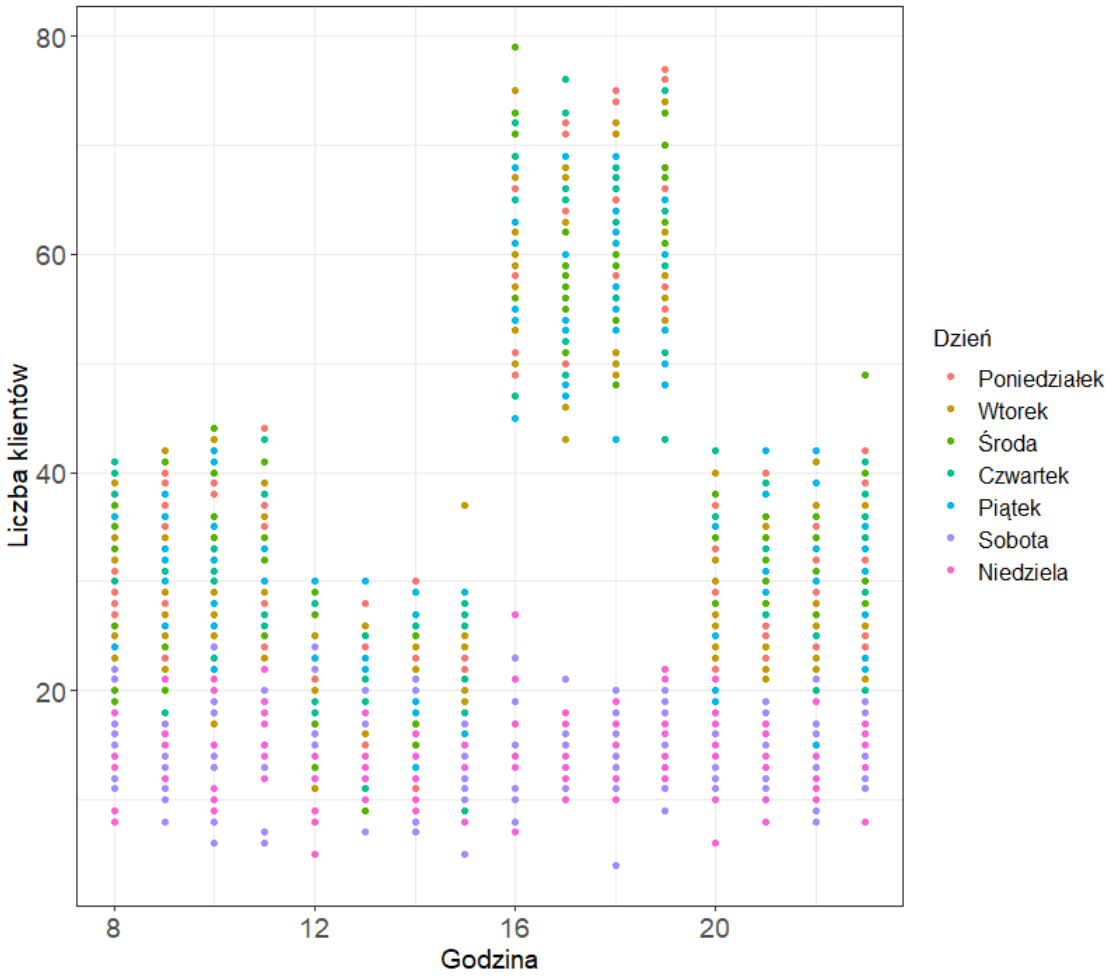
Również po spojrzeniu na dane pod kątem dnia tygodnia i godziny możemy zauważyć pewne prawidłowości, które opisałem już wcześniej.

Spróbujemy spojrzeć jeszcze na inne wykresy, które może dodatkowo wzmocnią nasze podejrzenia.

Jako, że podejrzewamy brak wpływu zmiennej *wydarzenie sportowe* na zmienną objaśnianą, spróbujmy, spojrzeć na wykresy bez rozbicia ze względu na *wydarzenie sportowe*.

Poniższe dwa wykresy pokazują rozrzut danych ze względu na dzień tygodnia i godzinę.





Widzimy zatem, że nasze podejrzenia mogą być prawdziwe. Liczba obsłużonych klientów wydaje się nie zależeć od tego jaki jest dzień - powszedni lub weekendowy.

Dodatkowo godziny układają się w pewne bloki czterogodzinne tzn. w pewnym przedziale czasowym liczba obsłużonych klientów wydaje się być podobna.

Natomiast w dni weekendowe widzimy, że tak liczba wygląda na niezależną od godziny.

Zadanie 3

W tym zadaniu stworzymy model regresji Poissona z interakcjami pomiędzy wszystimi regresorami dla naszych danych, gdzie zmienna objaśniana to liczba obsłużonych klientów, a pozostałe zmienne to regresory.

Ile zmiennych ma taki model?

Uwzględniając wszystkie interakcje nasz model ma 224 zmienne. Uwzględniamy wszystkie regresory, które mają odpowiednio 16, 7 i 2 stany, czyli liczba zmiennych to $16 \cdot 7 \cdot 2 = 224$. W modelu od regresora *wydarzenie sportowe* zależy 112 zmiennych, ponieważ dzielimy liczbę wszystkich zmiennych przez 2.

Czy zmienna *wydarzenie sportowe* jest istotna?

W tym celu skorzystamy z testu chi-kwadrat. Oznaczmy M_f jako model pełny, a M_r jako model bez zmiennej *wydarzenie sportowe*, dodatkowo oznaczmy I jako zbiór indeksów wektora parametrów zależnych od tej zmiennej.

Badamy następującą hipotezę:

$$H_0 : (\forall i \in I)\beta_i = 0 \quad \text{vs.} \quad H_1 : (\exists i \in I)\beta_i \neq 0.$$

Statystyka testowa ma postać:

$$\chi^2 = D(M_r) - D(M_f) = 116.13.$$

Przy hipotezie zerowej ma ona asymptotycznie rozkład χ^2 z 112 stopniami swobody. P-wartość tego testu wynosi 0.3755, czyli na poziomie istotności $\alpha = 0.05$ nie odrzucilibyśmy hipotezy zerowej. Nasze podejrzenia się potwierdziły, czyli możemy założyć że zmienna *wydarzenie sportowe* nie ma wpływu na liczbę klientów.

Czy interakcje są istotne? W celu zbadanie istotności interakcji wykonamy podobny test chi-kwadrat. Oznaczmy ponownie M_f jako model pełny, a M_r jako model bez interakcji oraz I jako zbiór indeksów wektora parametrów interakcji.

Badamy następującą hipotezę:

$$H_0 : (\forall i \in I)\beta_i = 0 \quad \text{vs.} \quad H_1 : (\exists i \in I)\beta_i \neq 0.$$

Statystyka testowa ma postać:

$$\chi^2 = D(M_r) - D(M_f) = 1051.57.$$

Ma ona asymptotycznie rozkład χ^2 z 201 stopniami swobody. P-wartość jest mniejsza od 2.2×10^{-16}. Oznacza to, że praktycznie na każdym sensownym poziomie istotności odrzucamy hipotezę zerową. Dlatego stwierdzamy, że interakcje są istotne w naszym modelu.

Zadanie 4

W porzednim zadaniu nasz model miał dość dużo zmiennych. Chcielibyśmy zredukować ich ilość. W poprzednim zadaniu tesując istotność zmiennej *wydarzenie sportowe*, dostaliśmy, że jest ona nieistotnoa. Stąd nie bedziemy jej uwzględniać w tym modelu. Dodatkowo w zadaniu drugim na podstawie wykresów wywnioskowaliśmy kilka zależności. Stworzymy więc dwie nowe zmienne. Pierwszą opisującą to czy dzień jest dniem roboczym czy weekendowym. Drugą grupującą godziny każdego dnia w bloki 4–o godzinne. Skonstruujemy model Poissona z interakcją pomiędzy nowymi zmiennymi traktując je jako faktory.

Nasz model prezentuje się następująco:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{1i} X_{4i} + \hat{\beta}_6 X_{2i} X_{4i} + \hat{\beta}_7 X_{3i} X_{4i},$$

gdzie zmienne przyjmują wartości:

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

Ile zmiennych ma nowy model?

Nowopowstałe zmienne mają odpowienio 4 i 2 stany, stąd uproszczony model ma 8 zmiennych.

Czy nowy model różni się statystycznie od modelu z poprzedniego zadania?

Ponownie skorzystamy w tym celu z testu chi-kwadrat. Oznaczmy nasze modele, jako M_3 model z poprzedniego zadania oraz M_4 jako model z tego zadania.

Testujemy hipotezę:

$$H_0 : \text{modele nie różnią się statystycznie} \quad \text{vs.} \quad H_1 : \text{modele różnią się.}$$

Statysytka wygląda następująco:

$$\chi^2 = D(M_4) - D(M_3) = 192.85.$$

Przy H_0 ma ona asymptotyczny rozkład χ^2 z 216 stopniami swobody, a p-wartość wynosi 0.8694. Stąd nie odrzucamy H_0.

Oznacza to, że model dużo prostszy i łatwiejszy do analizy daje nam prawie tą samą informację co najpełniejszy model. Dzięki temu będziemy mogli wykonać kolejne zadania i analizę stosując uproszczony model.

Zadanie 5

Chcielibyśmy stworzyć tabelkę w oparciu o model z zadania 4 składającą się z czterech wierszy. W pierwszym znajdzie się informacja o podgrupie. W drugim średnia liczba obsłużonych klientów. W trzecim i czwartym wierszu postać predyktora dla danej podgrupy oraz jego wartość. Rodzielimy tą tabelę na dwie w zależności od typu dnia.

Poniżej znajduje się tabela dla dni roboczych.

Dni robocze				
Godziny	08:00-11:59	12:00-15:59	16:00-19:59	20:00-23:59
Średnia liczba klientów	30	19.7	59.6	30.0
Postać predyktora	$\hat{\beta}_0 + \hat{\beta}_4$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_4 + \hat{\beta}_5$	$\hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_4 + \hat{\beta}_6$	$\hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_7$
Wartość predyktora	3.4	2.98	4.09	3.4

Natomiast poniżej dla dni weekendowych.

Dni weekendowe				
Godziny	08:00-11:59	12:00-15:59	16:00-19:59	20:00-23:59
Średnia liczba klientów	14.8	15	14.9	14.4
Postać predyktora	$\hat{\beta}_0$	$\hat{\beta}_0 + \hat{\beta}_1$	$\hat{\beta}_0 + \hat{\beta}_2$	$\hat{\beta}_0 + \hat{\beta}_3$
Wartość predyktora	2.69	2.7	2.7	2.67

Zadanie 6

Analiza wstępna i wyniki w powyższej tabeli sugerują, że w weekend klienci przychodzą z tą samą częstotliwością o różnych godzinach. Przetestujemy, czy predyktory liniowe odpowiadające podgrupom godzin weekendowych są takie same.

W tym celu skorzystamy z testu Walda.

Chcemy porównać trzy predyktory postaci:

$\eta_1 = \beta_0$,

$\eta_2 = \beta_0 + \beta_1$,

$\eta_3 = \beta_0 + \beta_2$,

$\eta_4 = \beta_0 + \beta_3$.

Testujemy hipotezę:

$$H_0 : \eta_1 = \eta_2 = \eta_3 = \eta_4 \quad \text{vs.} \quad H_1 : (\exists i \neq j)\eta_i \neq \eta_j.$$

Widzimy, że każdy predyktor ma β_0. Po odjęciu jej dostajemy hipotezę zerową:

$$\beta_1 = \beta_2 = \beta_3 = 0.$$

Statystyka testowa ma postać:

$$W = (L'\hat{\beta})'(L\text{Cov}(\hat{\beta})L)^{-1}(L'\hat{\beta}),$$

gdzie $\text{Cov}(\hat{\beta})$ to macierz kowariancji $\hat{\beta}$, a L' w naszym przypadku ma postać:

$$L' = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Statystyka ta ma rozkład chi-kwadrat z trzema stopniami swobodyi wynosi 1.383. Stąd p-wartość tego testu wynosi 0.7095. Widzimy, że jest ona dość duża. Czyli raczej nie odrzucilibyśmy hipotezy zerowej, więc widocznie nasze predyktory z dużym prawdopodobieństwem są równe. Po raz kolejny nasze podejrzenia się potwierdziły. Stąd możemy stwierdzić, że w weekend klienci przychodzą z takim samym natężeniem niezależnie od godziny.

Zadanie 7

Na podstawie wyników naszej analizy chcielibyśmy zaplanować optymalny grafik dla pracowników tego sklepu, zakładając, że pracownik w ciągu godziny jest w stanie obsłużyć około 20 klientów.

Do wypełnienia grafiku właściciel powinien zatrudnić 5 pracowników. Moja propozycja jest taka, aby podzielić pracowników na dwa typy: tygodniowych oraz weekendowych. Oznaczmy pracowników literami alfabetu:

tygodniowi: A, B, C, D

weekendowi: E, F

Przedstawię dwie tabele osobno dla dni roboczych oraz weekendowych.

Dni robocze				
Godziny	08:00-11:59	12:00-15:59	16:00-19:59	20:00-23:59
Średnia liczba klientów do obsłużenia na godzinę	30	20	60	30
Pracownicy na zmianie	A, F	A, D	B, C, D, E	B, C

Dni weekendowe				
Godziny	08:00-11:59	12:00-15:59	16:00-19:59	20:00-23:59
Średnia liczba klientów do obsłużenia na godzinę	15	15	15	14
Pracownicy na zmianie	F	F	E	E

Liczbę pracowników w danych blokach dopasowywałem głównie do średniej liczby klientów. W niektórych przypadkach jednak wiele obserwacji odbiegało znacznie od średniej, więc musiałem zwiększyć ich ilość, aby pracownicy byli w stanie obsłużyć klientów. Natomiast, gdy liczba klientów w danej godzinie przekroczyłaby liczbę, którą pracownicy są w stanie osbłużyć powinna się ona wyrównać z innymi godzinami.

Pracownicy tygodniowi pracują tylko w dni robocze, 8 godzin każdego dnia. Przedstawiłem tylko przykładowy grafik, godziny pracy mogą być zmieniane według preferencji właściciela lub pracowników np. co tydzień.

Natomiat pracownicy weekendowy jako jedyni pracują w weekendy. Dodatkowo pracują na pół etatu w dni robocze. Oczywiście tak jak w poprzedniej tabeli, przedstawiony układ pracowników jest przykładowy i może być zmieniony.

Sumując godziny pracy, pracownicy tygodniowi pracują 40 godzin tygodniowo, a pracownicy weekendowi 36 godzin.