Nocuń_Jarosław

Jarosław Nocuń

2024-04-22

```
chapter4 <- read.csv("D:/STUDIA DYSK D/Anaiza strukturalnych zbiorów
danych/rintro-chapter4.csv", header=TRUE, sep=",",dec=".")
```

Zestaw danych dla 1000 klientów detalicznego sprzedawcy, który sprzedaje produkty w sklepach stacjonarnych oraz online. Zbiór zawiera następujące dane: cust.id: ID klienta. age: Wiek klienta. credit.score: Punktacja kredytowej klienta. email: Czy klient ma adres e-mail. distance.to.store: Odległość klienta do najbliższego sklepu. online.visits: Liczba wizyt klienta na stronie internetowej w ciągu roku. online.trans: Liczba transakcji online klienta w ciągu roku. online.spend: Wydatki online klienta w ciągu roku. store.trans: Liczba transakcji w sklepie stacjonarnym klienta w ciągu roku. store.spend: Wydatki w sklepie stacjonarnym klienta w ciągu roku. sat.service: Ocena satysfakcji z obsługi w sklepie. sat.selection: Ocena satysfakcji z wyboru produktów w sklepie.

```
sum(complete.cases(chapter4))
```

```
## [1] 659
```

```
colSums(is.na(chapter4))
```

```
##           cust.id               age      credit.score             email
##                 0                 0                 0                 0
## distance.to.store     online.visits      online.trans      online.spend
##                 0                 0                 0                 0
##       store.trans       store.spend       sat.service     sat.selection
##                 0                 0               341               341
```

659/1000 zawartych rekordów jest pełnych. W pozostałych 341 rekordach brakuje danych dotyczących oceny satysfakcji z obsługi w sklepie oraz oceny satysfakcji z wyboru produktów w sklepie. Pomimo tych braków zbiór danych jest zdatny do analizy, szczególnie jeśli do analiz nie zostaną wykorzystane oceny satysfakcji obsługi w sklepie i satysfakcji z wyboru produktów w sklepie.

Do analizy zostaną wykorzystane następujące dane: 1. online.spend 2. store.spend 3. credit.score 4. distance.to.store

```
#Wyświetlenie liczby obserwacji odstających dla zmiennych wykorzystanych do
analiz
#zmienna online.spend

Q1 <- quantile(chapter4$online.spend, 0.25)
Q3 <- quantile(chapter4$online.spend, 0.75)
IQR <- Q3 - Q1
```

```r
lower_limit <- Q1 - 1.5 * IQR
upper_limit <- Q3 + 1.5 * IQR

outliers <- chapter4$online.spend[chapter4$online.spend < lower_limit |
chapter4$online.spend > upper_limit]
length(outliers)
```

```
## [1] 135
```

```r
#zmienna store.spend
Q1 <- quantile(chapter4$store.spend, 0.25)
Q3 <- quantile(chapter4$store.spend, 0.75)
IQR <- Q3 - Q1

lower_limit <- Q1 - 1.5 * IQR
upper_limit <- Q3 + 1.5 * IQR

outliers <- chapter4$store.spend[chapter4$store.spend < lower_limit |
chapter4$store.spend > upper_limit]
length(outliers)
```

```
## [1] 49
```

```r
#zmienna credit.score
Q1 <- quantile(chapter4$credit.score, 0.25)
Q3 <- quantile(chapter4$credit.score, 0.75)
IQR <- Q3 - Q1

lower_limit <- Q1 - 1.5 * IQR
upper_limit <- Q3 + 1.5 * IQR

outliers <- chapter4$credit.score[chapter4$credit.score < lower_limit |
chapter4$credit.score > upper_limit]
length(outliers)
```

```
## [1] 8
```

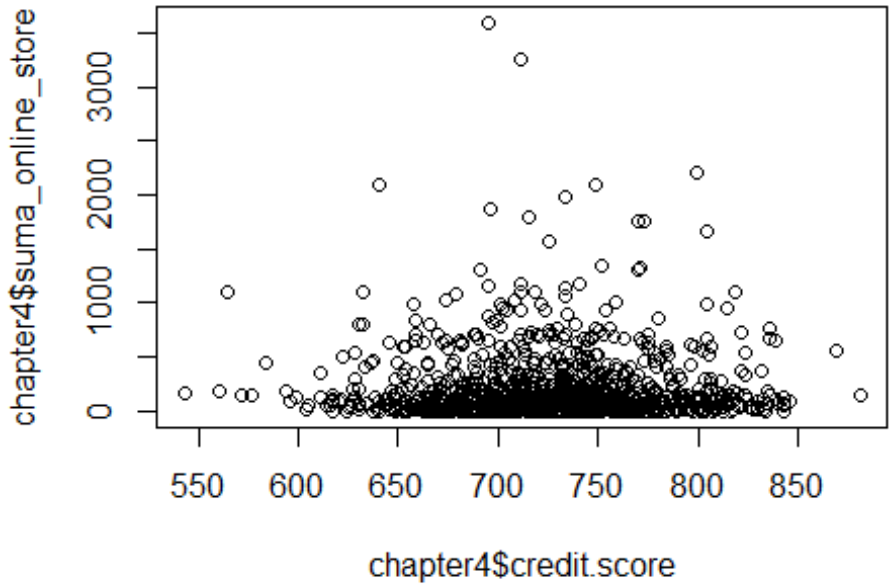
```r
#zmienna distance.to.store
Q1 <- quantile(chapter4$distance.to.store, 0.25)
Q3 <- quantile(chapter4$distance.to.store, 0.75)
IQR <- Q3 - Q1

lower_limit <- Q1 - 1.5 * IQR
upper_limit <- Q3 + 1.5 * IQR

outliers <- chapter4$distance.to.store[chapter4$distance.to.store <
lower_limit | chapter4$distance.to.store > upper_limit]
length(outliers)
```

```
## [1] 83
```

Przeanalizowano ilość obserwacji odstających dla zmiennych wykorzystanych do analiz. Dla zmiennej online.spend występuje ich 135, dla zmiennej store.spend 49, dla zmiennej credit.score 8 a dla zmiennej distance.to.store 83.

Czy zachodzi zależność pomiędzy credit.score a łączną kwotą wydaną online + fizycznie w sklepie przez klienta? H0: Zmienne X i Y są niezależne H1: Zmienne X i Y nie są niezależne

```
#Utworzenie nowej zmiennej sumującej zmienne onlie.spend oraz store.spend
chapter4$suma_online_store <- (chapter4$online.spend + chapter4$store.spend)

plot(chapter4$credit.score, chapter4$suma_online_store)
```




```
#Przeprowadzenie testu korelacji liniowej Pearsona
cor.test(chapter4$suma_online_store, chapter4$credit.score, method="pearson")

##
##  Pearson's product-moment correlation
##
## data:  chapter4$suma_online_store and chapter4$credit.score
## t = 0.084065, df = 998, p-value = 0.933
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.05934186  0.06464343
## sample estimates:
##         cor
## 0.002661011
```

Wartość p jest zdecydowanie większa niż 0,05. Nie ma podstaw do odrzucenia hipotezy zerowej, więc zmienne credit.score i łączna kwota wydana online + fizycznie w sklepie przez klienta są niezależne.

Czy średnie wydatki online są równe dla klientów z różną odległością do sklepu?

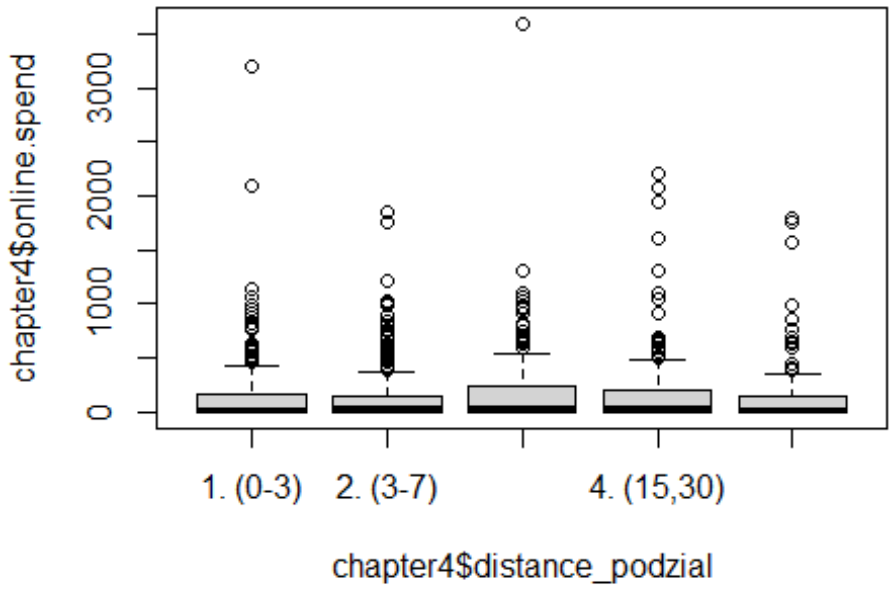
By porównać średnie wydatki online z odległością do sklepu należy najpierw odległość podzielić na grupy. Zmienna ta została podzielona na 5 grup: grupa 1: od 0 do 3 grupa 2: od 3 do 7 grupa 3: od 7 do 15 grupa 4: od 15 do 30 grupa 5: od 30 do 267

```
#Podzielenie zmiennej distance.to.store na 5 grup
chapter4$distance_podzial <- cut(chapter4$distance.to.store, br=c(0, 3, 7,
15, 30, 280), labels =c("1. (0-3)", "2. (3-7)", "3. (7,15)", "4. (15,30)",
"5. (30,267)"), na.rm=TRUE)

table(chapter4$distance_podzial)

##
##    1. (0-3)    2. (3-7)   3. (7,15)  4. (15,30) 5. (30,267)
##         211         283         219         165         122

#Histogram zmiennej distance.to.store
hist(chapter4$distance.to.store)
```

Z histogramu odległości do sklepu można zauważyć, że klienci z grupy 5 (czyli ci, których odległośc do sklepu jest większa niż 30) stanowią zdecydowaną mniejszość.

Następnym prezentowanym wykresem jest wykres pudełkowy, który może dostarczyć informacji dotyczących rozkładu zmiennych.

```
#Wykres pudełkowy dla zmiennych online.spend oraz distance_podzial
boxplot(chapter4$online.spend~chapter4$distance_podzial)
```



Następnie przeprowadzony został test Shapiro-Wilka dla poszczególnych grup. H0: średnie wydatki online są równe dla klientów z różnymi odległościami do sklepu H1: Istnieje przynajmniej jedna para klientów z różną odległością do sklepu dla których średnie wydatki online są różne

```
#Test Shapiro-Wilka
tapply(chapter4$online.spend,chapter4$distance_podzial,shapiro.test)
```

```
## $`1. (0-3)`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.5186, p-value < 2.2e-16
##
##
## $`2. (3-7)`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.61517, p-value < 2.2e-16
```

```
##
##
## $`3. (7,15)`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.55684, p-value < 2.2e-16
##
##
## $`4. (15,30)`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.56315, p-value < 2.2e-16
##
##
## $`5. (30,267)`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.52548, p-value < 2.2e-16
```

P-value jest mniejsze od 0,05 we wszystkich przypadkach, dlatego odrzucamy hipotezę zerową na korzyść hipotezy alternatywnej. Rozkłady nie są normalne.

Test Kruskala-Wallisa: H0: rozkład wydatków online w każdej grupie odległości od sklepu jest równy H1: rozkład wydatków online nie w każdej grupie odległości od sklepu jest równy

```
#Test Kruskala-Wallisa
kruskal.test(chapter4$online.spend~chapter4$distance_podzial)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  chapter4$online.spend by chapter4$distance_podzial
## Kruskal-Wallis chi-squared = 2.6441, df = 4, p-value = 0.619
```

Wartość p jest większa niż 0,05. Brak podstaw do odrzucenia H0. Rozkład wydatków online w każdej grupie odległości od sklepu jest równy.