Regresja liniowa prosta

Contents

Opis danych **1**

Dopasowanie oraz analiza modeli **13**
Regresja liniowa dla price 14
Regresja liniowa dla advertising 20
Regresja liniowa dla age 21

Podsumowanie **22**

```
carseats <- tibble::as.tibble(ISLR::Carseats)
head(carseats)
```

```
## # A tibble: 6 x 11
##   Sales CompPrice Income Advertising Population Price ShelveLoc   Age Education
##   <dbl>     <dbl>  <dbl>       <dbl>      <dbl> <dbl> <fct>     <dbl>     <dbl>
## 1  9.5        138     73          11        276   120 Bad          42        17
## 2 11.2        111     48          16        260    83 Good         65        10
## 3 10.1        113     35          10        269    80 Medium       59        12
## 4  7.4        117    100           4        466    97 Medium       55        14
## 5  4.15       141     64           3        340   128 Bad          38        13
## 6 10.8        124    113          13        501    72 Bad          78        16
## # i 2 more variables: Urban <fct>, US <fct>
```

```
dim(carseats)
```

```
## [1] 400  11
```

Opis danych

Zaczniemy od opisania naszego zbioru danych. Zbiór ten posiada 400 wierszy oraz 11 kolumn które dotyczą sprzedaży fotelików samochodowych. Opiszemy co oznaczają poszczególne kolumny.
Sales: Sprzedaż fotelików w tysiącach jednostek.
CompPrice: Cena konkurencyjnego produktu w danym regionie.
Income: Średni dochód w regionie (w tysiącach dolarów).
Advertising: Budżet reklamowy w danym regionie (w tysiącach dolarów).
Population: Populacja w regionie (w tysiącach).
Price: Cena fotelika.
ShelveLoc: Lokalizacja półki z fotelikami (kategorie: *Bad*, *Good*, *Medium*).
Age: Średni wiek mieszkańców w regionie.
Education: Średni poziom edukacji w regionie.
Urban: Zmienna wskazująca, czy region jest miejski (*Yes*) lub wiejski (*No*).
US: Zmienna wskazująca, czy region znajduje się w USA (*Yes*) lub poza USA (*No*).
Nastepnie policzymy podstawowe statystyki dla naszych danych.

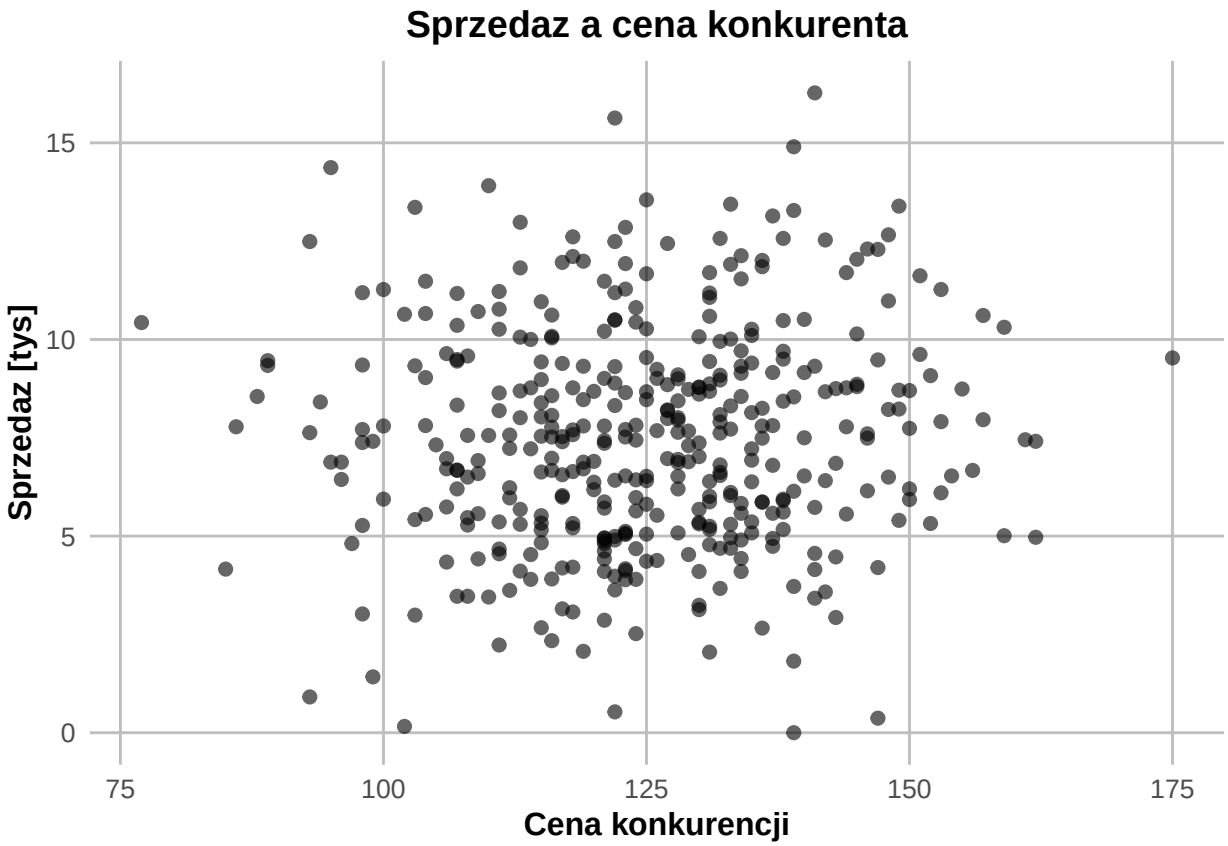
```
summary(carseats)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
##  Median : 7.490   Median :125   Median : 69.00   Median : 5.000
##  Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
##    Population        Price        ShelveLoc        Age          Education
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0
##  Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0
##  Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0
##  Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0
##  Urban       US
##  No :118   No :142
##  Yes:282   Yes:258
##
##
##
##
```

Zbadamy zależności między zmiennymi objaśniającymi, a zmienną Sales poprzez wykresy punktowe oraz obliczenie korelacji Pearsona.

```
ggplot(carseats, aes(x = CompPrice, y = Sales)) + geom_point(size = 2, color = "black", alpha = 0.6) +
  labs(title = 'Sprzedaż a cena konkurenta', x = 'Cena konkurencji', y = 'Sprzedaż [tys]') +  theme_min
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5),
    panel.grid.minor = element_blank()
  )
```

```
## Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
## i Please use the `linewidth` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

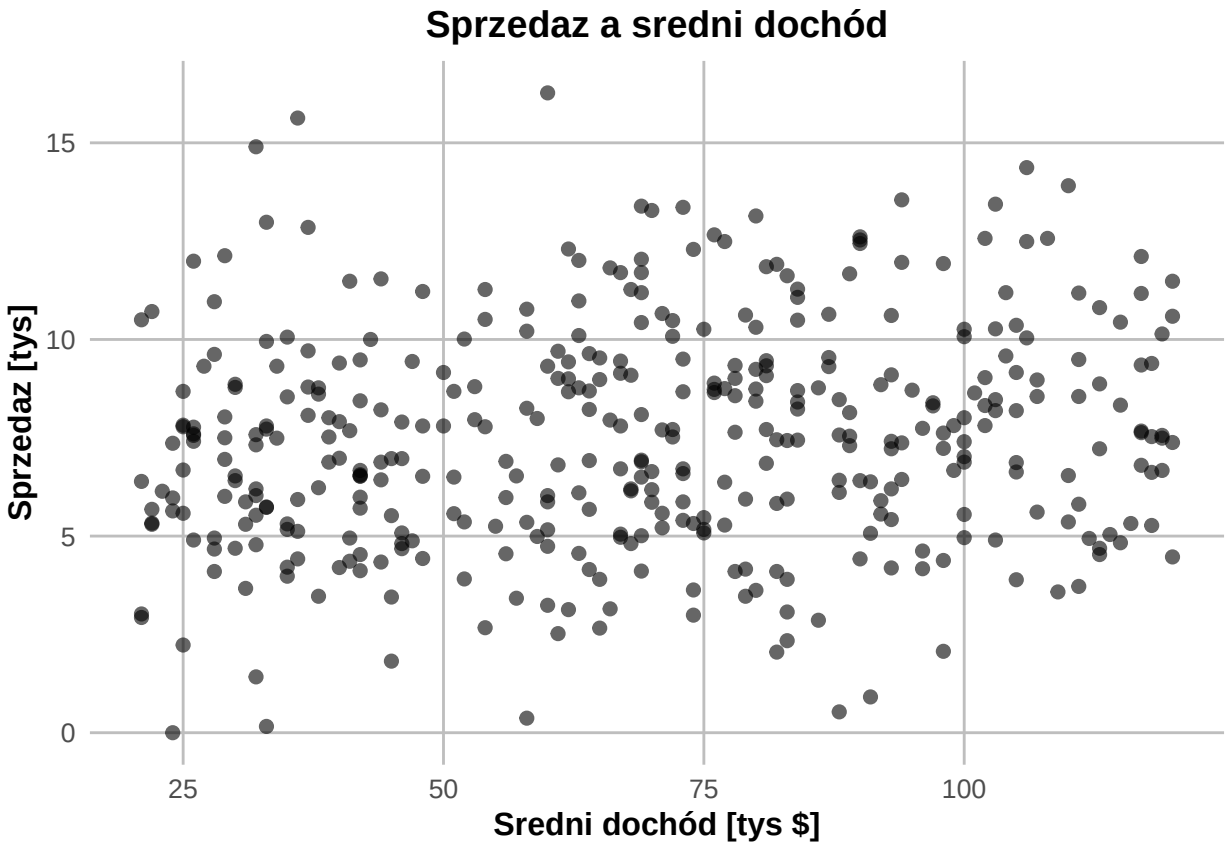



```
cor(carseats$CompPrice, carseats$Sales)
```

```
## [1] 0.06407873
```

Możemy zauważyć, że istnieje bardzo znikoma korelacja między sprzedażą a ceną u konkurencji. Co może sugerować, że cena konkurencji nie ma istotnego wpływu na sprzedaż w naszym przypadku.

```
ggplot(carseats, aes(x = Income, y = Sales)) + geom_point(size = 2, color = "black", alpha = 0.6) +
  labs(title = 'Sprzedaż a średni dochód', x = 'Średni dochód [tys $]', y = 'Sprzedaż [tys]') + theme_m
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5),
    panel.grid.minor = element_blank()
  )
```

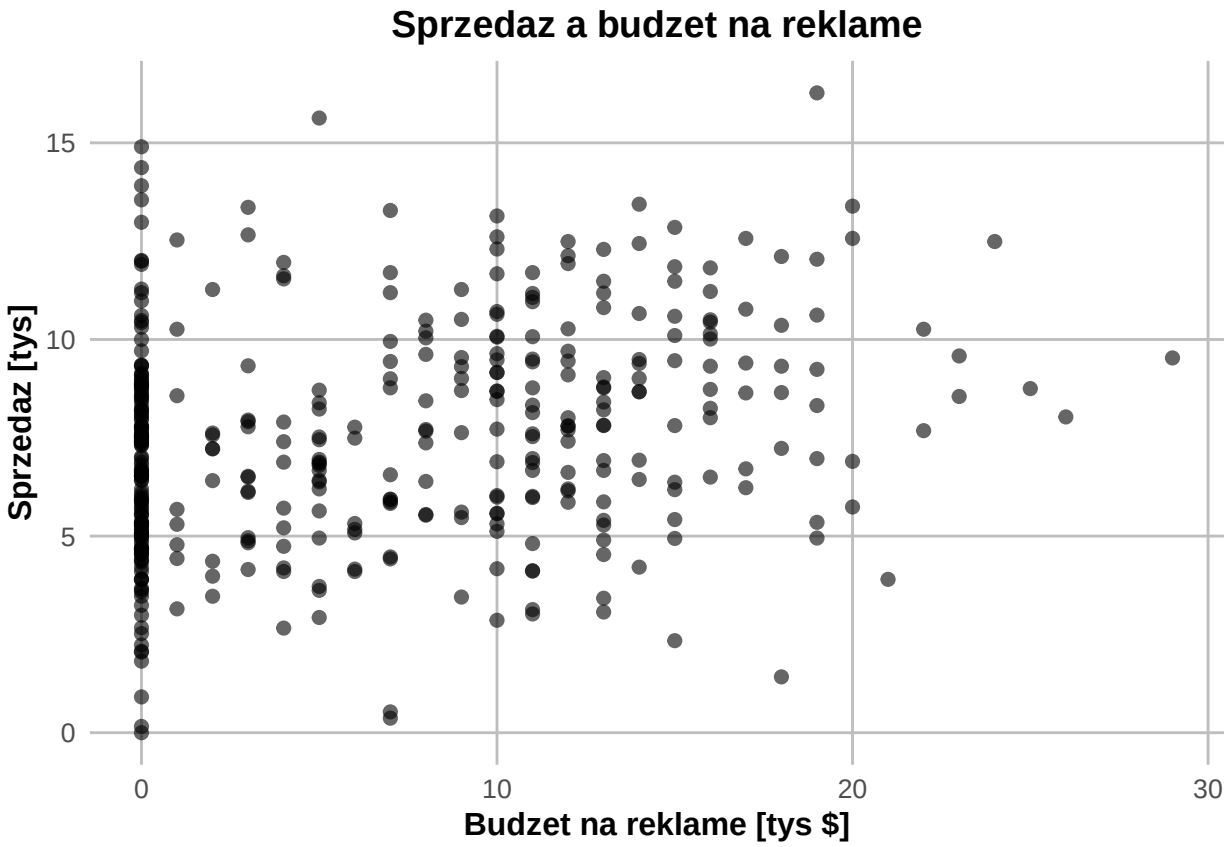



```
cor(carseats$Income, carseats$Sales)
```

```
## [1] 0.151951
```

Sprzedaż oraz dochód w danym obszarze wykazują bardzo niską korelację dodatnią. Sugeruje to, że związek między zmiennymi jest znikomy.

```
ggplot(carseats, aes(x = Advertising, y = Sales)) + geom_point(size = 2, color = "black", alpha = 0.6) +
  labs(title = 'Sprzedaż a budżet na reklame', x = 'Budżet na reklame [tys $]', y = 'Sprzedaż [tys]') +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5),
    panel.grid.minor = element_blank()
  )
```

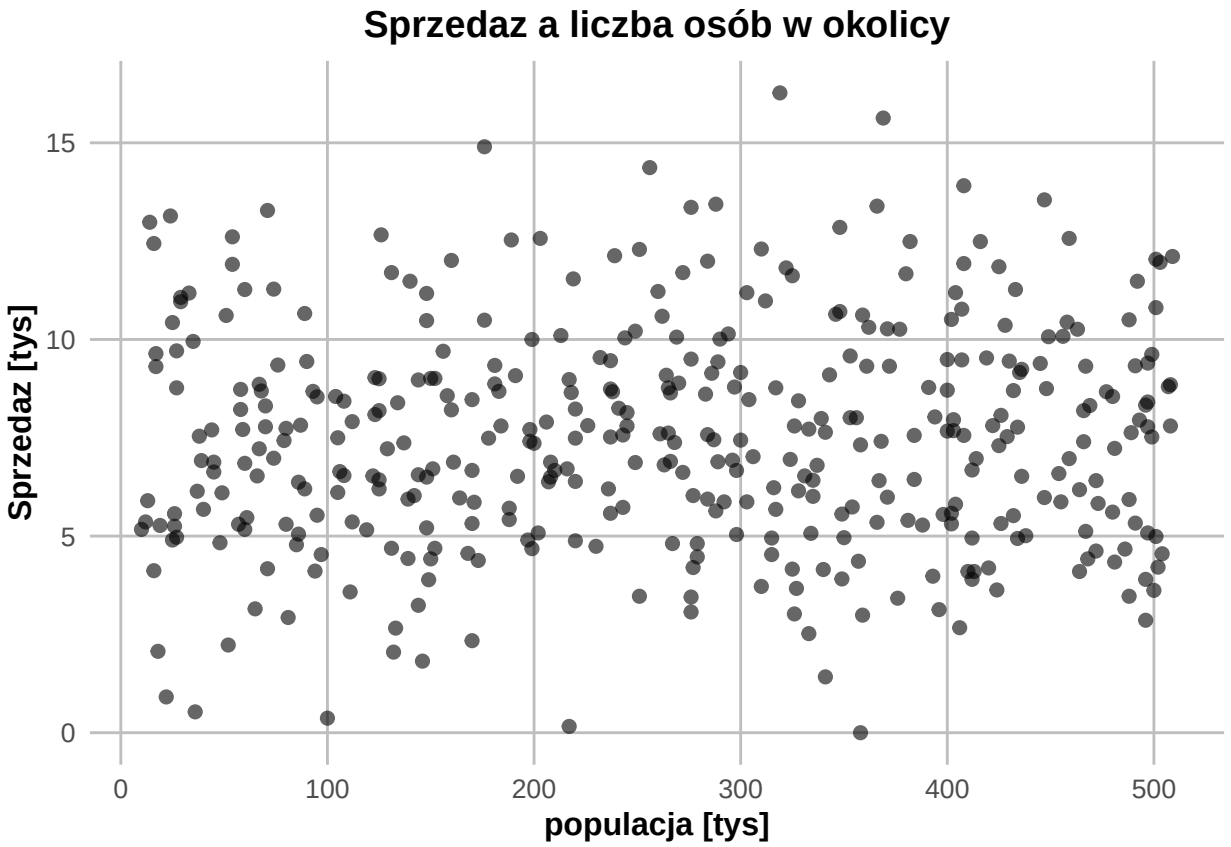



```
cor(carseats$Advertising, carseats$Sales)
```

```
## [1] 0.2695068
```

Budżet reklamowy oraz sprzedaż wykazują wyraźniejszy związek niż dwie powyższe lecz nie jest on silny. Może wskazywać na to, że wieksze wydatki na reklamę mogą stymulować popyt.

```
ggplot(carseats, aes(x = Population, y = Sales)) + geom_point(size = 2, color = "black", alpha = 0.6) +
  labs(title = 'Sprzedaż a liczba osób w okolicy', x = 'populacja [tys]', y = 'Sprzedaż [tys]') + theme_
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5),
    panel.grid.minor = element_blank()
  )
```

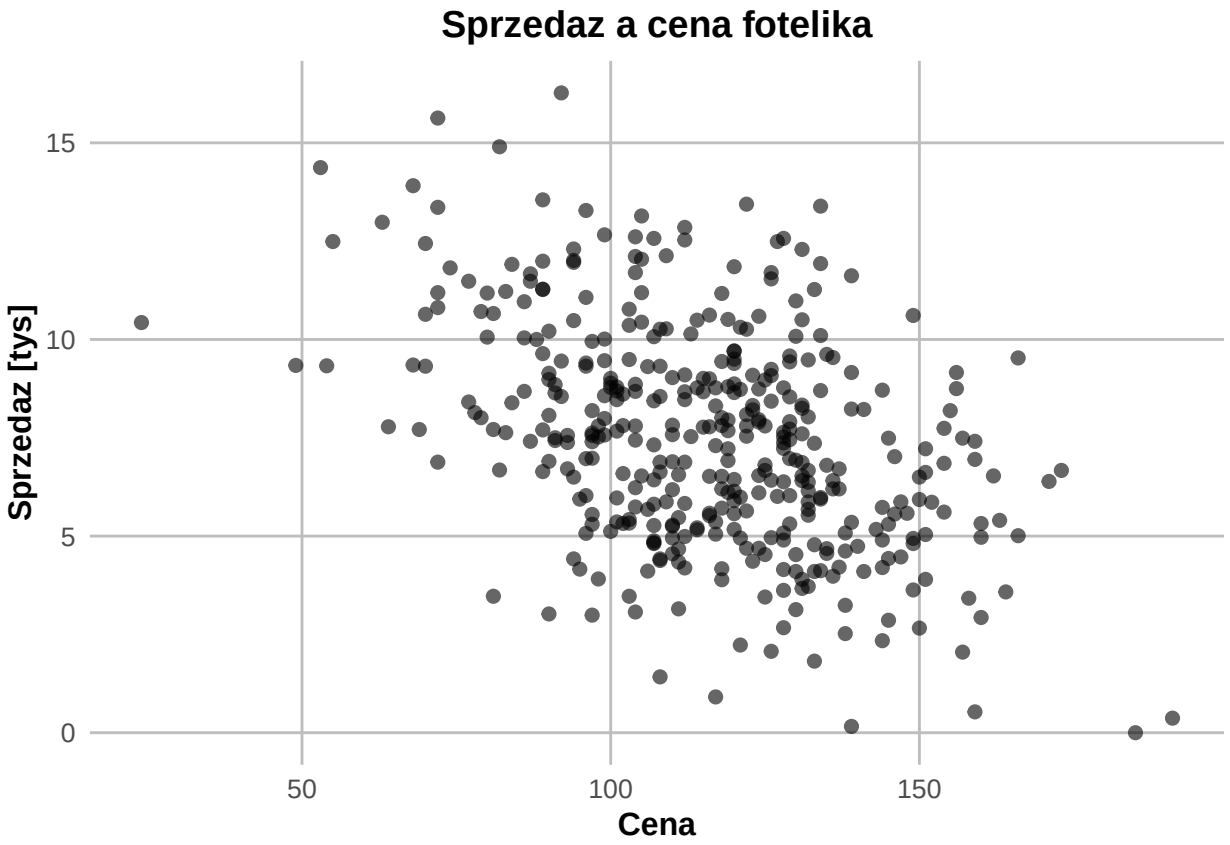



```
cor(carseats$Population, carseats$Sales)
```

```
## [1] 0.05047098
```

Wielkosc populacji w danym regionie nie ma istotnego wpływu na sprzedaż.

```
ggplot(carseats, aes(x = Price, y = Sales)) + geom_point(size = 2, color = "black", alpha = 0.6) +
  labs(title = 'Sprzedaż a cena fotelika', x = 'Cena', y = 'Sprzedaż [tys]') + theme_minimal(base_size =
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5),
    panel.grid.minor = element_blank()
  )
```

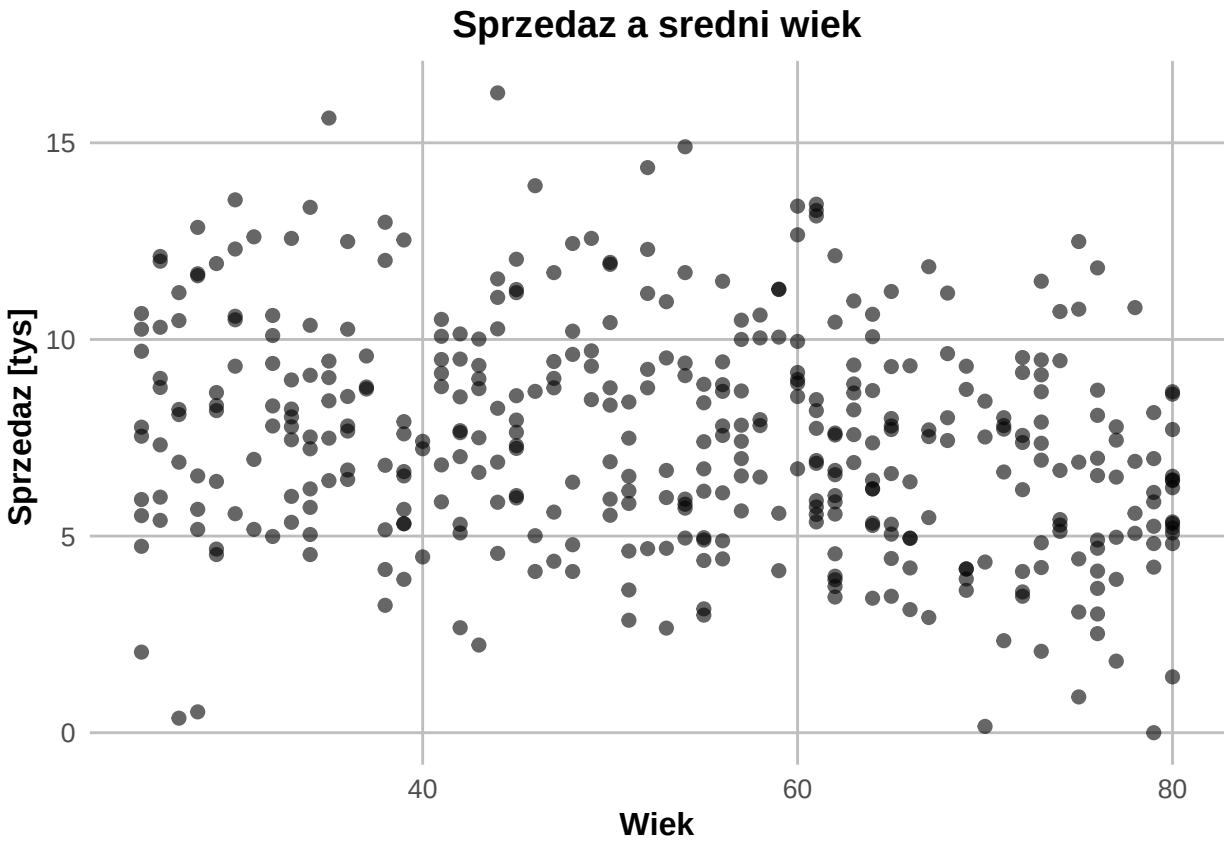



```
cor(carseats$Price, carseats$Sales)
```

```
## [1] -0.4449507
```

Cena fotelika oraz sprzedaż wykazują ujemna korelacje. Co sugeruje, że wraz ze wzrostem ceny sprzedaż może mieć tendencje do spadków.

```
ggplot(carseats, aes(x = Age, y = Sales)) + geom_point(size = 2, color = "black", alpha = 0.6) +
  labs(title = 'Sprzedaż a średni wiek', x = 'Wiek', y = 'Sprzedaż [tys] ') + theme_minimal(base_size =
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5),
    panel.grid.minor = element_blank()
  )
```

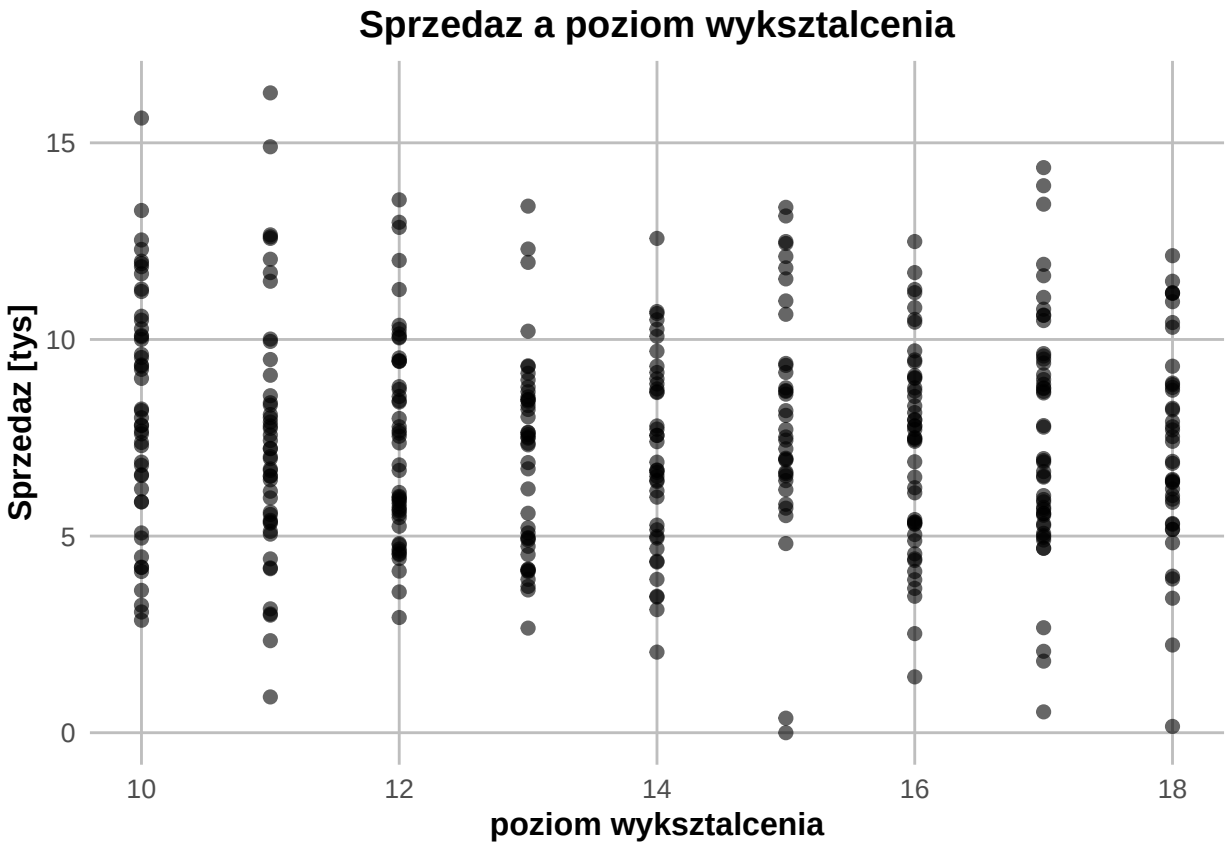



```
cor(carseats$Age, carseats$Sales)
```

```
## [1] -0.2318154
```

średni wiek mieszkańców w regionie oraz sprzedaż wykazują słabą ujemną korelację, co może oznaczać, że na obszarach z wyższym średnim wiekiem może wystąpic mniejsza sprzedaż.

```
ggplot(carseats, aes(x = Education, y = Sales)) + geom_point(size = 2, color = "black", alpha = 0.6) +
  labs(title = 'Sprzedaż a poziom wykształcenia', x = 'poziom wykształcenia', y = 'Sprzedaż [tys]') + tl
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5),
    panel.grid.minor = element_blank()
  )
```

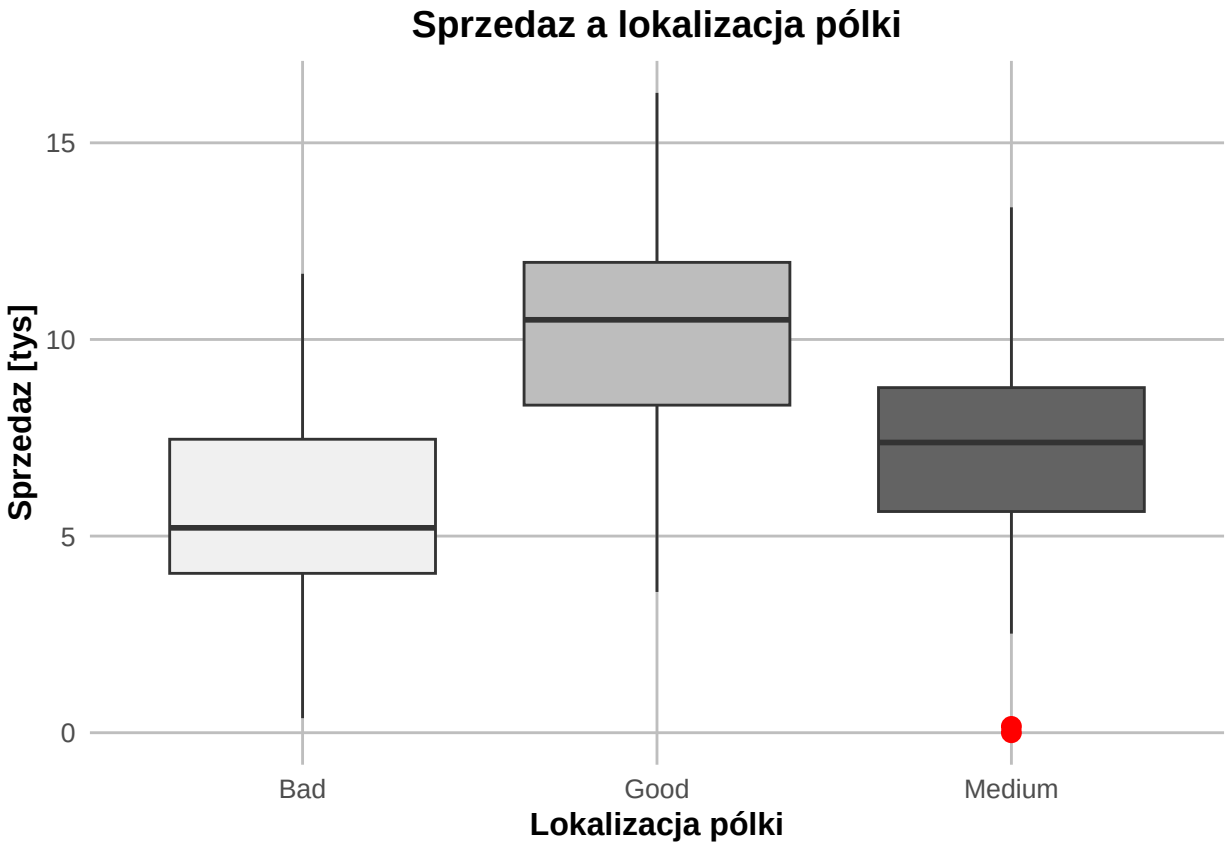



```
cor(carseats$Education, carseats$Sales)
```

```
## [1] -0.05195524
```

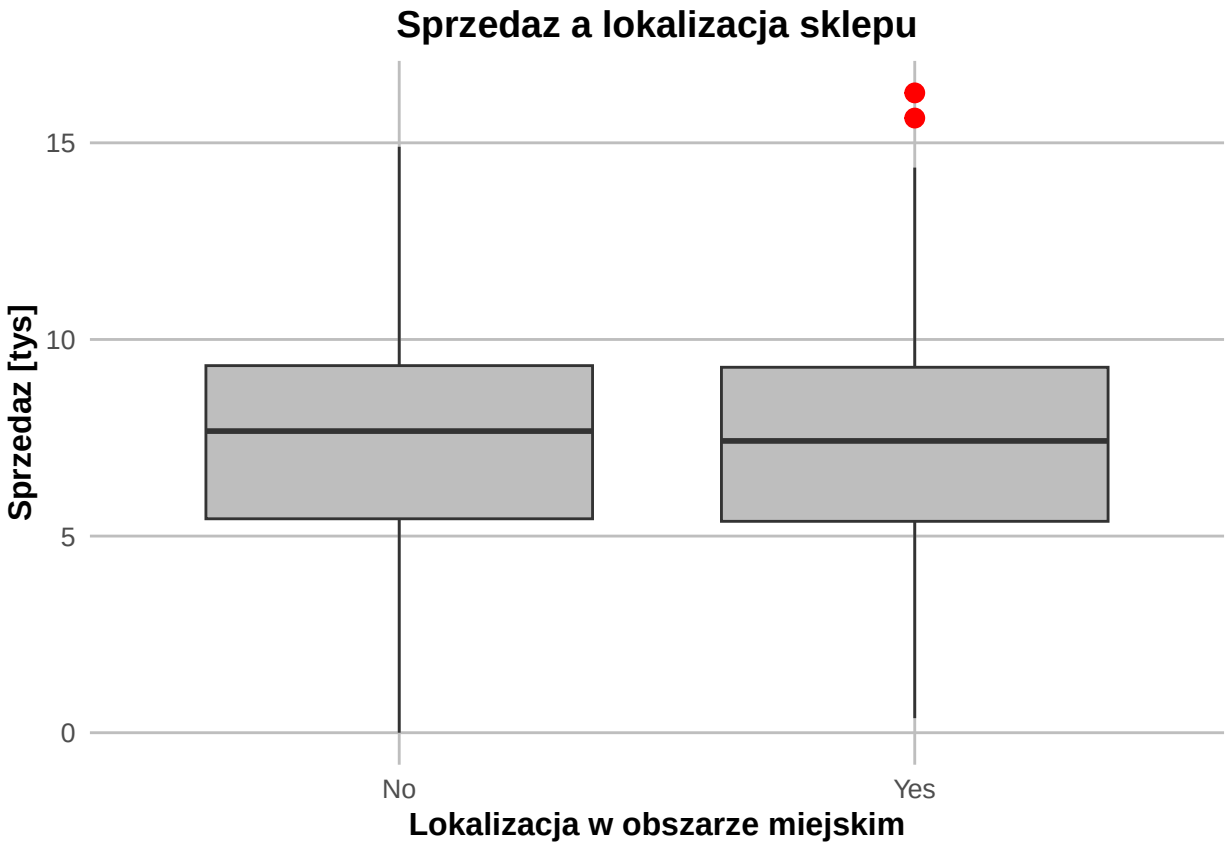
Możemy zauważyć ze zmienne wykazują bardzo znikomą korelację ujemną wskazując na to, że poziom wykształcenia nie wpływa na sprzedaż.
Zobaczymy jak prezentują się zmienne kategorialne.

```
ggplot(Carseats, aes(x=ShelveLoc, y=Sales, fill=ShelveLoc)) + geom_boxplot(outlier.color='red', outlier
  scale_fill_brewer(palette = "Greys") +
  labs(title = "Sprzedaż a lokalizacja półki", x = "Lokalizacja pólki", y = "Sprzedaż [tys]") +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5),
    panel.grid.minor = element_blank(),
    legend.position = "none"
  )
```



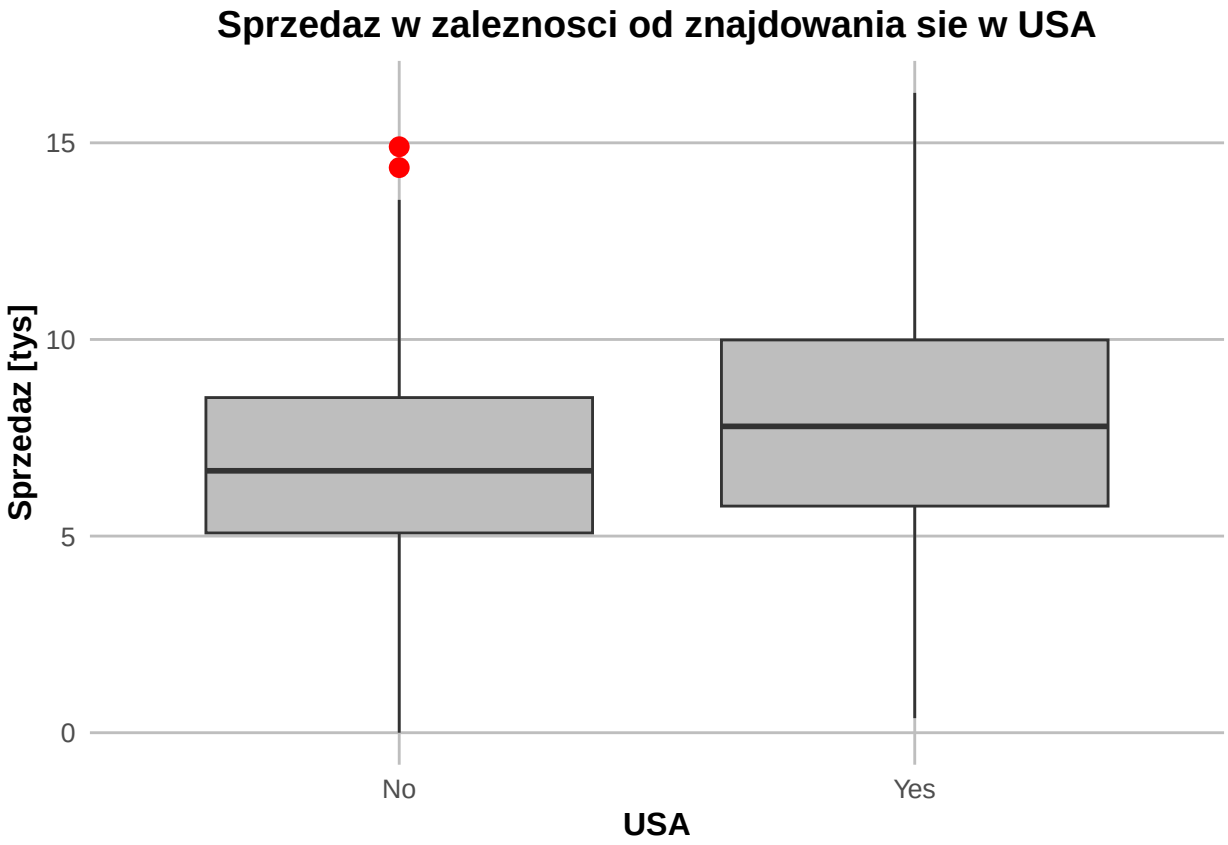
Możemy zauważyć, że coraz lepsza lokalizacja półki w sklepie wykazuje najlepszą sprzedaż.

```
ggplot(Carseats, aes(x=Urban, y=Sales)) + geom_boxplot(fill='grey', outlier.color='red', outlier.size=3)
  labs(title = "Sprzedaż a lokalizacja sklepu", x = "Lokalizacja w obszarze miejskim", y = "Sprzedaż [ty
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5),
    panel.grid.minor = element_blank()
  )
```



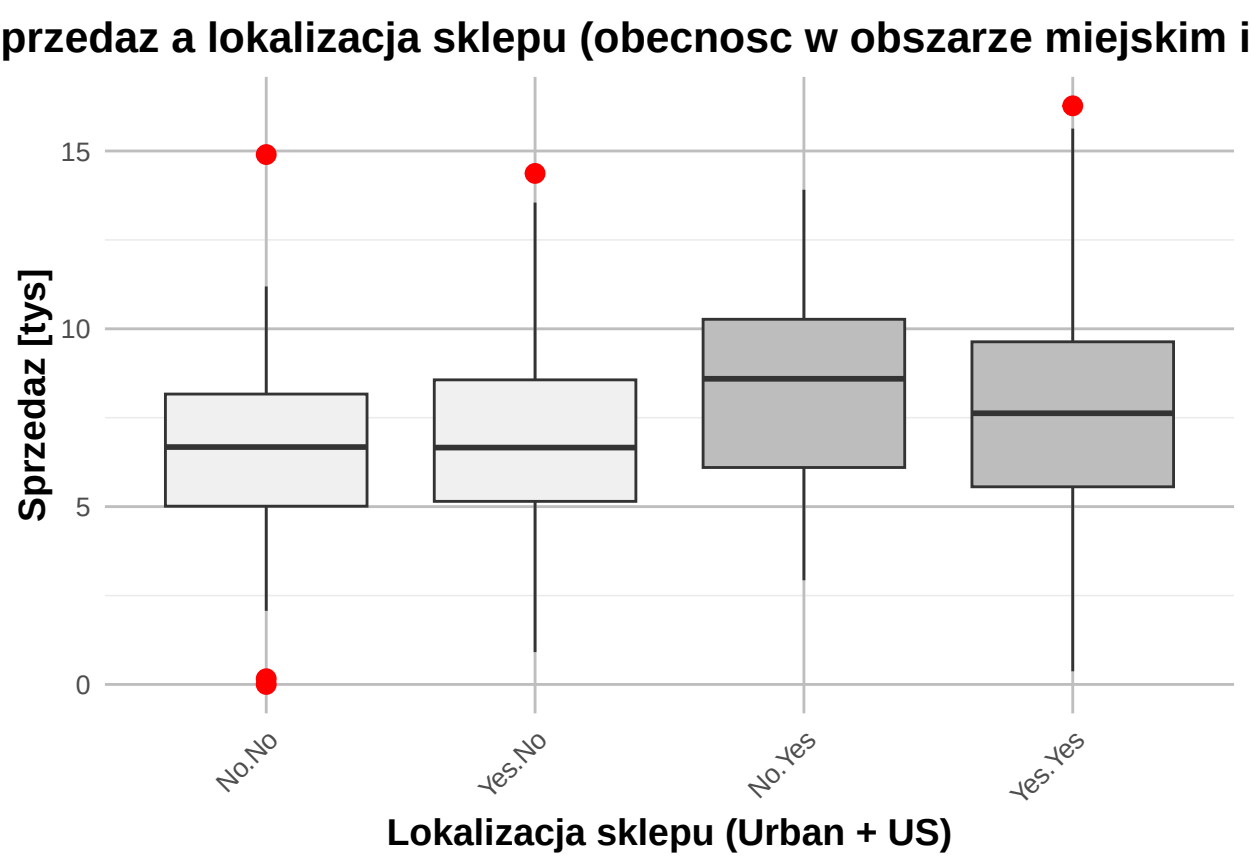
Znikoma różnica między sklepami znajdującymi się w mieście a sklepami poza miastem.

```
ggplot(Carseats, aes(x=US, y=Sales)) + geom_boxplot(fill='grey', outlier.color='red', outlier.size=3) +
  labs(title="Sprzedaż w zależności od znajdowania się w USA", x='USA', y='Sprzedaż [tys]') +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5),
    panel.grid.minor = element_blank()
  )
```



Sklepy znajdujące się w USA mają minimalnie wyższą sprzedaż

```
ggplot(Carseats, aes(x = interaction(Urban, US), y = Sales, fill = US)) +
  geom_boxplot(outlier.color = 'red', outlier.size = 3) +
  scale_fill_brewer(palette = "Greys") +
  labs(title = "Sprzedaż a lokalizacja sklepu (obecność w obszarze miejskim i USA)",
       x = "Lokalizacja sklepu (Urban + US)",
       y = "Sprzedaż [tys]") +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    axis.title = element_text(face = "bold", size = 14),
    panel.grid.major = element_line(color = "gray", size = 0.5),
    legend.position = "none",
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



Możemy zauważyć, że sprzedaż sklepów znajdujących się w USA jest znacznie wieksza niż sklepów nie znajdujących się w USA. Również róznica występuje gdy sklep znajduję sie w USA ale nie jest w obszasze miejskim, wiekszą mediane sprzedaży obserwujemy w przypadku gdy sklep nie znajduje się w obszarze miejskim.
Podsumowując wstepne zapoznanie się ze zbiorem danych:

```
cor(Carseats %>% select_if(is.numeric))["Sales", -which(colnames(Carseats) == "Sales")]
```

```
##   CompPrice      Income Advertising  Population       Price         Age
##  0.06407873  0.15195098  0.26950678  0.05047098 -0.44495073 -0.23181544
##   Education
## -0.05195524
```

Zwróćmy uwagę na to, że korelacja zmiennej Sales z zmiennymi CompPrice, Population i Education jest bliska 0, więc już teraz możemy stwierdzić, że modele regresji liniowej prostej z tymi zmiennymi objaśniającymi nie będą miały dobrej jakości.

Dopasowanie oraz analiza modeli

Dopasujemy teraz modele regresji liniowej przewidujące wartość zmiennej Sales oraz zbadamy poniższe założenia:

1. Zależność liniowa między zmiennymi
2. Rozkład reszt mający rozkład normalny
3. Zerowa średnia reszt
4. Niezależność reszt
5. Homoskedastyczność stała wariancja błędów

```r
#podzielenie zbioru na zbiór treningowy i zbiór testowy
set.seed(44)
partition <- caret::createDataPartition(carseats$Sales, list=FALSE, p=0.75)
carseats_train <- carseats[partition,]
carseats_test <- carseats[-partition,]
```

```r
print(dim(carseats_train))
```

```
## [1] 301  11
```

```r
print(dim(carseats_test))
```

```
## [1] 99 11
```

```r
MAPE <- function(y_actual, y_predicted) {
  y_actual[y_actual == 0] <- NA  # Obsługa potencjalnych zer w mianowniku
  return(mean(abs((y_actual - y_predicted) / y_actual), na.rm = TRUE) * 100)
}
RMSE <- function(y_actual, y_predicted){
  return(sqrt(mean((y_actual-y_predicted)^2)))
}
R_2 <- function(y_actual, y_predicted) {

  ss_total <- sum((y_actual - mean(y_actual))^2)
  ss_residual <- sum((y_actual - y_predicted)^2)

  r_squared <- 1 - (ss_residual / ss_total)

  return(r_squared)
}
```

Regresja liniowa dla price

```r
price_model <- lm(Sales ~ Price, data=carseats_train)
```

Sprawdzenie zależości liniowej:

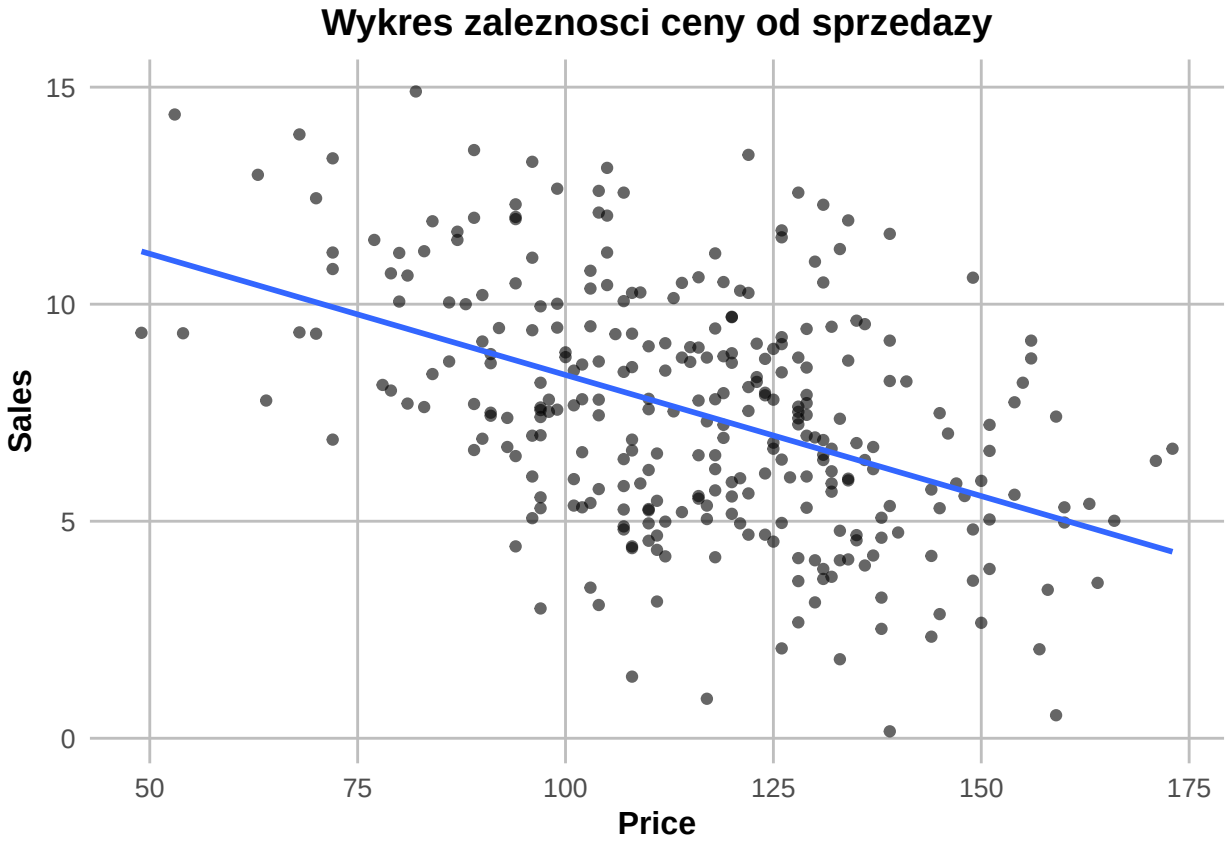
```r
cor.test(carseats_train$Price, carseats_train$Sales)
```

```
##
##  Pearson's product-moment correlation
##
## data:  carseats_train$Price and carseats_train$Sales
## t = -8.8095, df = 299, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5393225 -0.3593370
## sample estimates:
##        cor
## -0.4539481
```

```r
ggplot(carseats_train, aes(x=Price, y=Sales)) + geom_point(size = 1.5, color = "black", alpha = 0.6) +
  geom_smooth(method='lm' ,formula=y~x, size=1, se=FALSE) +
  labs(title='Wykres zależności ceny od sprzedaży') + theme_minimal(base_size = 12) +
  theme(
```

```
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5),
    panel.grid.minor = element_blank()
  )
```

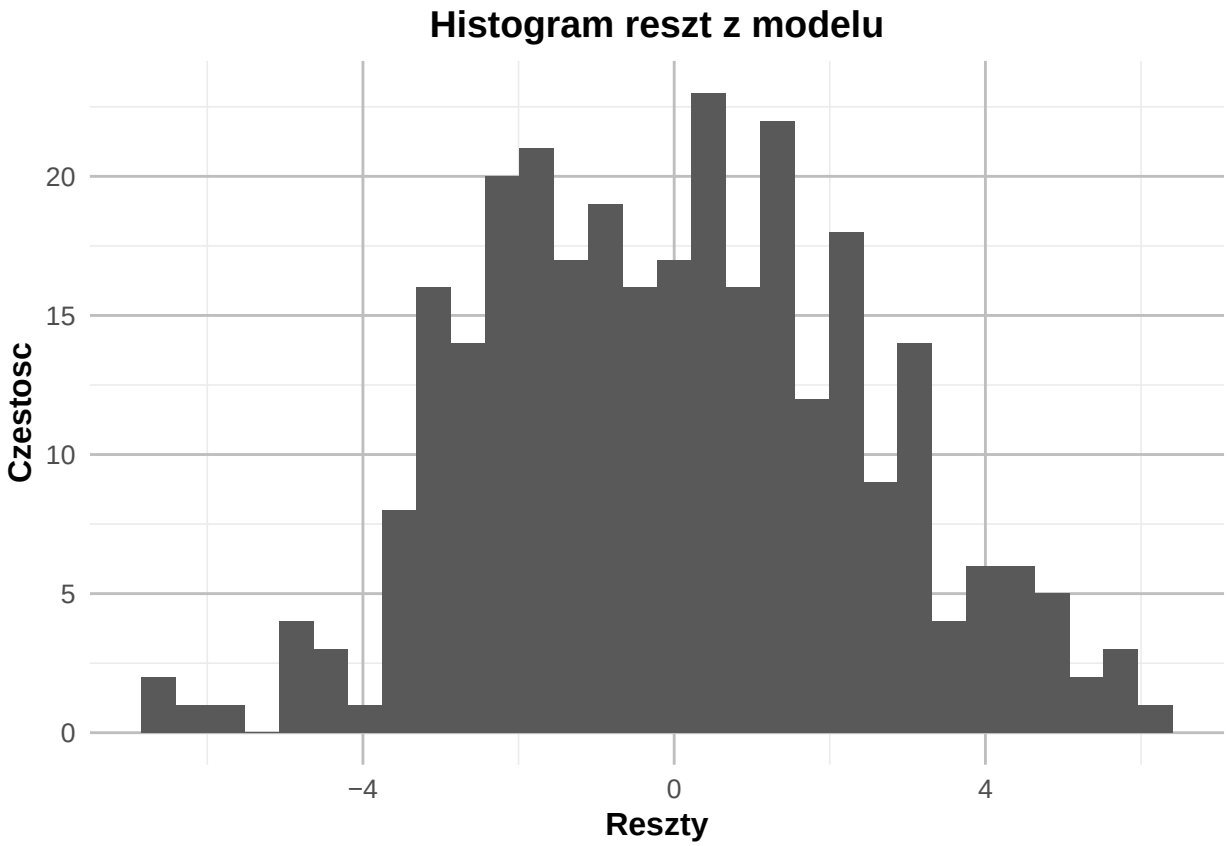
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



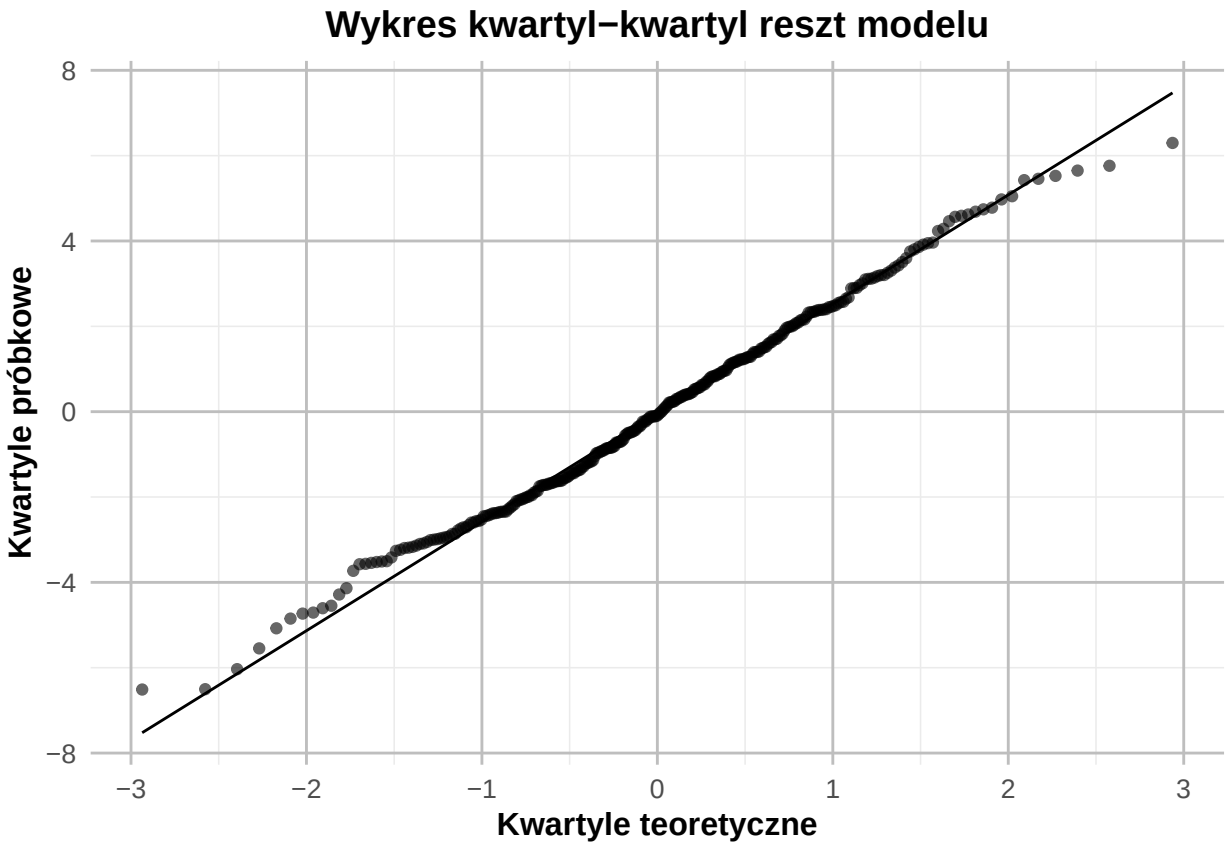
Korelacja oraz wykres wskazują na negatywną korelację.

Sprawdzenie rozkładu reszt:

```
ggplot(price_model, aes(x=.resid)) + geom_histogram(bins=30) +
  labs(title='Histogram reszt z modelu', x='Reszty', y='Częstość') + theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```




```
ggplot(price_model, aes(sample = .resid)) +
  geom_qq(size = 1.5, color = "black", alpha = 0.6) +
  geom_qq_line() +
  labs(title = 'Wykres kwartyl-kwartyl reszt modelu',
       x = 'Kwartyle teoretyczne',
       y = 'Kwartyle próbkowe') +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```




```
shapiro.test(price_model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  price_model$residuals
## W = 0.99407, p-value = 0.289
```

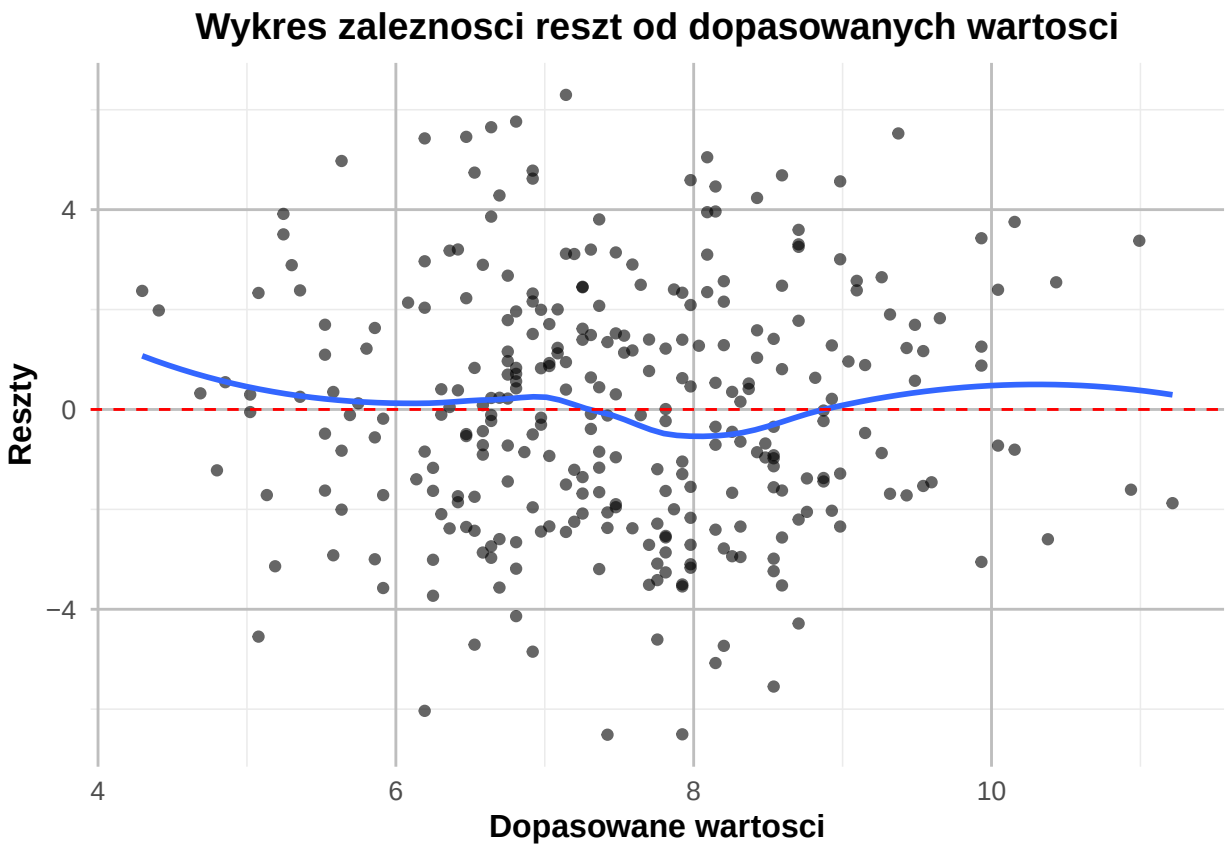
Powyższcze wykresy oraz test wskazują, że większość reszt odpowiada rozkładowi normalnemu. ### Sprawdzenie zerowej średniej reszt:

```
t.test(price_model$residuals)
```

```
##
##  One Sample t-test
##
## data:  price_model$residuals
## t = -4.9495e-16, df = 300, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.2799931  0.2799931
## sample estimates:
##     mean of x
## -7.042199e-17
```

Test wykazał, że należy odrzucić hipotezę alternatywną oraz możemy stwierdzić, że średnia reszt jest równa zero.

```
ggplot(price_model, aes(.fitted, .resid)) + geom_point(size = 1.5, color = "black", alpha = 0.6) + stat_
  geom_hline(yintercept=0, linetype='dashed', color='red') +
  labs(title='Wykres zależności reszt od dopasowanych wartości', x='Dopasowane wartości',y='Reszty') + t
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```



Sprawdzenie niezależności reszt:

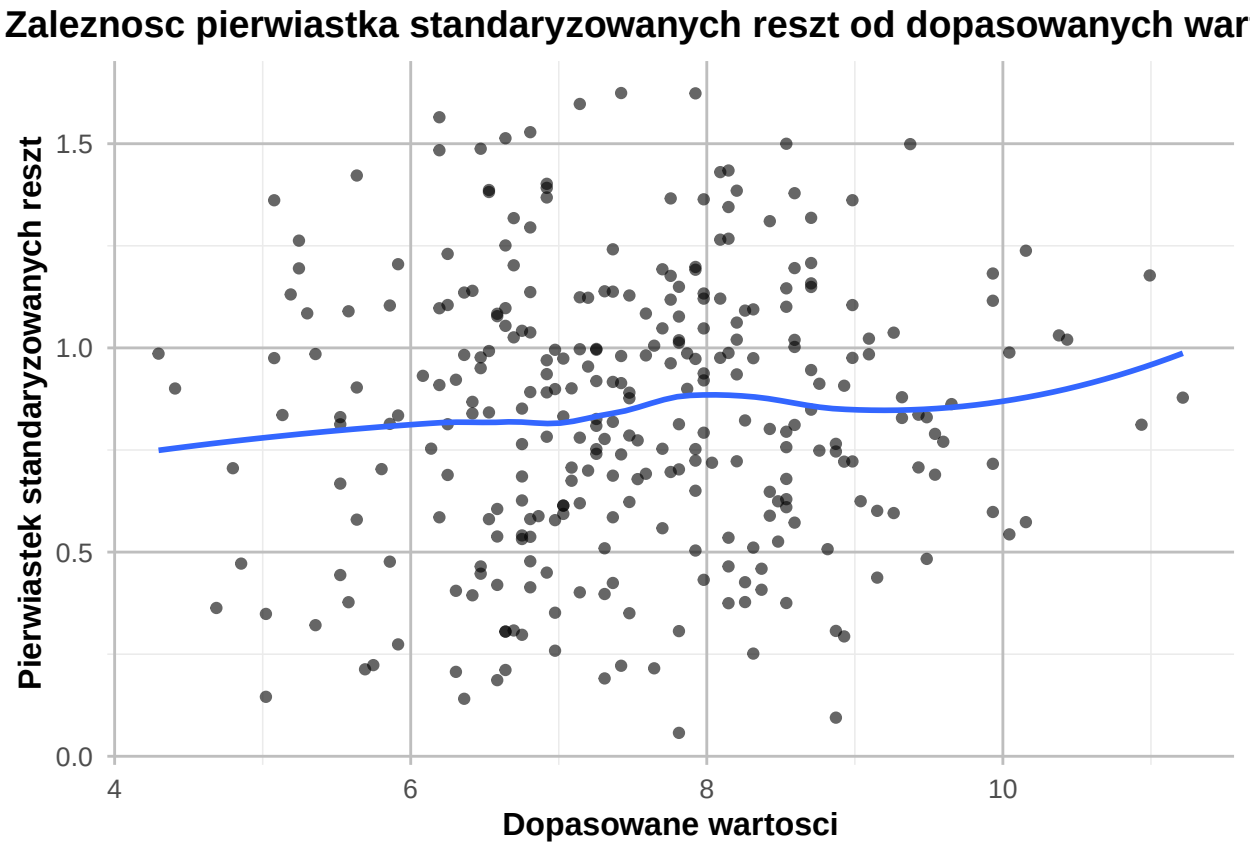
```
lmtest::dwtest(price_model)
```

```
##
##  Durbin-Watson test
##
## data:  price_model
## DW = 1.9414, p-value = 0.3046
## alternative hypothesis: true autocorrelation is greater than 0
```

Przeprowadzony test wykazał, że p-value jest większe od $\alpha = 0{,}05$ więc nie mamy dowodów, aby odrzucić hipotezę o niezależności reszt. Wartośc DW jest bliska 2 co wskazuje na brak autokorelacji.Możemy wnioskować,że założenie o niezależności reszt jest spełnione dla naszego modelu.

Sprawdzenie homoskedatyczność:

```
ggplot(price_model, aes(.fitted, sqrt(abs(.stdresid)))) + geom_point(size = 1.5, color = "black", alpha
  labs(title='Zależność pierwiastka standaryzowanych reszt od dopasowanych wartości', x='Dopasowane war
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```




```
lmtest::bptest(price_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  price_model
## BP = 0.019352, df = 1, p-value = 0.8894
```

Wartość p-value wyniosłą znacznie więcej niż $\alpha = 0{,}05$, oznacza to, że nie mamy istotnych dowodów heteroskedastyczności. Dlatego też możemy wnioskować,że założenie o homoskedastyczności jest prawdziwe dla naszego modelu.

Podsumowanie modelu price:

Nasz model spełnia klasyczne założenia modelu regresji.

```
price_model_s <- summary(price_model)
price_model_s
```

```
##
## Call:
## lm(formula = Sales ~ Price, data = carseats_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5113 -1.7485 -0.0897  1.6958  6.2977
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.949676   0.747060  18.673   <2e-16 ***
## Price       -0.055798   0.006334  -8.809   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.473 on 299 degrees of freedom
## Multiple R-squared:  0.2061, Adjusted R-squared:  0.2034
## F-statistic: 77.61 on 1 and 299 DF,  p-value: < 2.2e-16
```

```
test_predictions <- predict(price_model, carseats_test)
```

```
cat('Zbiór treningowy:', '\n-MAPE:',MAPE(carseats_train$Sales, predict(price_model, carseats_train)), '
```

```
## Zbiór treningowy:
## -MAPE: 50.51755
## -RMSE: 2.464361
## -R^2: 0.2060689
```

```
cat('Zbiór testowy:', '\n-MAPE:',MAPE(carseats_test$Sales, test_predictions), '\n-RMSE:', RMSE(carseats_
```

```
## Zbiór testowy:
## -MAPE: 41.02037
## -RMSE: 2.70789
## -R^2: 0.1748445
```

Regresja liniowa dla advertising

zakladamy ze model spelnia zalozenia

```
adv_model <- lm(Sales ~ Advertising, data=carseats_train)
```

```
adv_model_s <- summary(adv_model)
adv_model_s
```

```
##
## Call:
## lm(formula = Sales ~ Advertising, data = carseats_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2080 -1.8992 -0.0792  1.7814  8.0108
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.88919    0.21814  31.582  < 2e-16 ***
## Advertising  0.09660    0.02455   3.935 0.000104 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.706 on 299 degrees of freedom
## Multiple R-squared:  0.04924,    Adjusted R-squared:  0.04606
## F-statistic: 15.48 on 1 and 299 DF,  p-value: 0.0001035
```

```
test_predictions <- predict(adv_model, carseats_test)
```

```
cat('Zbiór treningowy:', '\n-MAPE:',MAPE(carseats_train$Sales, predict(adv_model, carseats_train)), '\n
```

```
## Zbiór treningowy:
## -MAPE: 56.36055
## -RMSE: 2.696803
## -R^2: 0.04923643
```

```
cat('Zbiór testowy:', '\n-MAPE:',MAPE(carseats_test$Sales, test_predictions), '\n-RMSE:', RMSE(carseats_
```

```
## Zbiór testowy:
## -MAPE: 52.55149
## -RMSE: 2.785452
## -R^2: 0.1268976
```

Model wykazuje lepsze wyniki niż poprzedni model, choć nadal jego dokładność jest ograniczona. W zbiorze treningowym, MAPE wynosi 56,36%, co wskazuje na średni błąd prognozy wynoszący 56%, a RMSE to 2,70, co oznacza, że model wciąż nie jest precyzyjny. Wartość R^2 w zbiorze treningowym wynosi 0,049, co wskazuje, że model wyjaśnia tylko około 5% zmienności sprzedaży, co nadal jest stosunkowo niskim wynikiem.

Zbiór testowy wykazuje nieco lepsze wyniki w porównaniu do zbioru treningowego: MAPE wynosi 52,55%, RMSE to 2,79, a R^2 wynosi 0,127. Chociaż R^2 na zbiorze testowym jest nieco wyższe, nadal oznacza to, że model nie wyjaśnia większej części zmienności sprzedaży.

Mimo że zmienna Advertising ma statystycznie istotny wpływ na sprzedaż, jakość modelu pozostaje niewystarczająca, aby uznać go za precyzyjny.

Regresja liniowa dla age

zakladamy ze model spelnia zalozenia

```
age_model <- lm(Sales ~ Age, data=carseats_train)
```

```
age_model_s <- summary(age_model)
age_model_s
```

```
##
## Call:
## lm(formula = Sales ~ Age, data = carseats_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9898 -1.8832 -0.0326  1.7696  7.4346
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.655286   0.531491  18.166  < 2e-16 ***
## Age         -0.040554   0.009517  -4.261 2.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.694 on 299 degrees of freedom
## Multiple R-squared:  0.05725,    Adjusted R-squared:  0.0541
## F-statistic: 18.16 on 1 and 299 DF,  p-value: 2.728e-05
```

```
test_predictions <- predict(age_model, carseats_test)
```

```
cat('Zbiór treningowy:', '\n-MAPE:',MAPE(carseats_train$Sales, predict(age_model, carseats_train)), '\n
```

```
## Zbiór treningowy:
## -MAPE: 56.57889
## -RMSE: 2.685411
## -R^2: 0.05725182
```

```
cat('Zbiór testowy:', '\n-MAPE:',MAPE(carseats_test$Sales, test_predictions), '\n-RMSE:', RMSE(carseats
```

```
## Zbiór testowy:
## -MAPE: 55.77424
## -RMSE: 2.913969
## -R^2: 0.04447128
```

W zbiorze treningowym MAPE wynosi 56,58%, co oznacza średni błąd prognozy na poziomie około 57%. RMSE wynosi 2,69, a R^2 osiąga wartość 0,057, co oznacza, że model wyjaśnia tylko około 5,7% zmienności w danych. Zmienna Age jest statystycznie istotna, co sugeruje, że wpływa na prognozowaną sprzedaż, ale jej wpływ jest ograniczony.

Zbiór testowy pokazuje marginalne pogorszenie wyników: MAPE wynosi 55,77%, RMSE osiąga wartość 2,91, a R^2 spada do 0,044, co oznacza, że model nie poprawia swojej wydajności na nowych danych.

Ogólnie, chociaż model wskazuje na pewną istotność zmiennej Age, jakość prognoz jest słaba, z niskim R^2 i wysokimi błędami prognozy. Model nie spełnia oczekiwań w kontekście precyzyjności przewidywań.

Podsumowanie

Model oparty na zmiennej **price** (cena) wydaje się być najlepszym modelem w tym przypadku. Uzyskał najniższy **MAPE** w zbiorze testowym (41,02%) oraz stosunkowo wysokie wartości R^2 (0,1748) w porównaniu do innych zmiennych.