Eksploracja i analiza danych z katastrofy Titanica

Cel:

Celem niniejszego sprawozdania jest przeprowadzenie analizy danych dotyczących pasażerów statku Titanic, zawartych w pliku titanic_new.csv. Analiza ma na celu zrozumienie struktury danych, wykonanie odpowiedniego wstępnego przetwarzania (preprocessingu), przeprowadzenie podstawowej analizy statystycznej oraz identyfikację potencjalnych zależności między zmiennymi.

Dzięki odpowiedniemu przetwarzaniu oraz eksploracji danych możliwe będzie sformułowanie hipotez analitycznych, np. które czynniki mogły wpływać na przeżycie pasażerów. Taka analiza może stanowić podstawę do budowy modelu predykcyjnego klasyfikującego, czy dany pasażer miał większe szanse na przeżycie.

Teoria:

Zbiór titanic_new.csv zawiera informacje o pasażerach Titanica, w tym dane demograficzne, dane o bilecie i opłacie, miejscu zaokrętowania, numerze kabiny, a także informację, czy pasażer przeżył katastrofę.

Preprocessing (czyli wstępne przetwarzanie danych) to etap przygotowania danych przed analizą statystyczną lub budową modelu predykcyjnego. To **pierwszy i kluczowy krok** w pracy z danymi, który ma na celu uczynienie ich czystymi, spójnymi i gotowymi do dalszej analizy.

Po co robimy preprocessing?

- Usuwa błędy i niespójności
- Poprawia jakość danych
- Zapobiega błędom w analizach i modelach
- Ułatwia interpretację wyników
- Przygotowuje dane do wykorzystania w algorytmach statystycznych lub machine learning.

Statystyki opisowe
To podstawowe miary używane do podsumowania i opisania cech danych liczbowych. Obejmują m.in. średnią, medianę, min/max, odchylenie standardowe i kwartyle. Pomagają szybko zrozumieć rozkład i zakres zmiennych.

Test chi-kwadrat (χ^2)
To test statystyczny służący do sprawdzania, czy istnieje zależność między dwiema zmiennymi kategorycznymi. Porównuje wartości obserwowane z oczekiwanymi. Jeśli różnice są duże, sugeruje to istotną zależność.

Korelacja
Miara siły i kierunku związku między dwiema zmiennymi liczbowymi. Przyjmuje wartości od -1 (silna korelacja ujemna) do 1 (silna dodatnia). Korelacja bliska 0 wskazuje na brak liniowego związku między zmiennymi.

Preprocessing:

1. Czyszczenie danych:

Zostały usunięte następujące kolumny:

- **Passenger.Id**
  To unikalny numer identyfikacyjny, który nie niesie żadnej informacji, mogącej pomóc przewidzieć np. przeżycie. Kolumna ta nie jest użyteczna w analizie ani modelowaniu.
- **Name**
  Imię i nazwisko pasażera są unikalnymi danymi (każde imię jest inne), więc wprowadzałyby „szum". Bez dodatkowego przetwarzania (np. analiza tytułów lub długości imienia) nie niosą one żadnej przydatnej informacji.
- **Ticket**
  Bilet może zawierać literę lub numer, ale jego znaczenie nie jest jasne. W praktyce rzadko pomaga w modelowaniu, a często zawiera dużo niejednorodnych danych tekstowych.
- **Cabin**
  Wiele wierszy ma wartość NA (brak danych), a same dane są niejednorodne i często niekompletne. Ze względu na duży odsetek braków i brak znaczącej wartości informacyjnej, kolumna ta została usunięta.

2. Transformacja danych:

Kolumny **Survived**, **Pclass**, **Sex**, **Embarked** zostały zamienione na dane kategoryczne (typ factor):

- **Survived**
  To zmienna kategoryczna, a nie liczbowa (1 = przeżył, 0 = nie przeżył). Nie chcemy, by model traktował 1 jako "więcej" niż 0. Zamiana na factor pozwala analizować przeżycie jako klasy, co jest bardziej odpowiednie w analizach klasyfikacyjnych.
- **Pclass**
  Oznacza klasę biletu (1, 2, 3). Choć zapisane są jako liczby, klasy są kategoriami, a nie wartościami liczbowymi. Zamiana na factor pozwala na traktowanie ich jako kategorie i umożliwia właściwą analizę.
- **Sex**
  Zmienna kategoryczna (mężczyzna, kobieta). Zamiana na factor umożliwia porównanie grup mężczyzn i kobiet w analizie.
- **Embarked**
  Miejsce zaokrętowania pasażera (C, Q, S). Zamiana na factor pozwala analizować zależności między miejscem zaokrętowania a innymi zmiennymi, np. przeżyciem.

3. Uzupełnienie brakujących danych:

- **Age** – uzupełnienie **średnią**
  Wiek (Age) to zmienna ciągła numeryczna. W przypadku brakujących danych, możemy uzupełnić je średnią.Dzięki temu nie tracimy danych, a zmienna pozostaje gotowa do analizy.

- **Embarked** – uzupełnienie **najczęściej występującą wartością** (modą)
  W przypadku zmiennych kategorycznych, takich jak Embarked (port zaokrętowania), brakujące dane najlepiej uzupełnić najczęściej występującą kategorią. Dzięki temu zachowujemy spójność typów danych i nie zakłócamy rozkładu.

4. Tworzenie nowych zmiennych:

- **FamilySize = SibSp + Parch + 1**
  Kolumny **SibSp** (liczba rodzeństwa i małżonków) oraz **Parch** (liczba rodziców i dzieci) zostały połączone, tworząc zmienną **FamilySize**. Dodanie 1 uwzględnia samego pasażera. Tworzenie tej zmiennej pozwala na lepsze zrozumienie wpływu wielkości rodziny na przeżycie pasażera. Zmienna ta dzieli pasażerów na kategorie: samotnych, z małą rodziną, z dużą rodziną. Zmienna FamilySize została również wykorzystana w dalszej analizie korelacji oraz statystyk opisowych.

Część analityczna:

Celem analizy jest poznanie struktury danych, rozkładów zmiennych oraz relacji między nimi, w szczególności tych, które mogą mieć wpływ na przeżycie pasażerów (Survived).

1. Statystyki opisowe zmiennych liczbowych

Przeanalizowano zmienne liczbowe: Age, Fare, SibSp, Parch, FamilySize. Obliczono ich średnie, mediany, minimalne i maksymalne wartości.

Statystyka	Age	SibSp	Parch	FamilySize
Min	0.42	0	0	1
1st Qu.	22.00	0	0	1
Median	29.70	0	0	1
Mean	29.70	0.523	0.3816	1.905
3rd Qu.	35.00	1	0	2
Max	80.00	8	6	11

1. **Wiek (Age)**

- Wiek pasażerów waha się od **0.42** do **80** lat.
- **Średni wiek** wynosi około **29.7 lat**, a **mediana** również 29.7, co sugeruje symetryczny rozkład.
- Pierwszy kwartyl (Q1) to 22 lata, a trzeci (Q3) – 35 lat, co oznacza, że większość pasażerów miała od 22 do 35 lat.
- Obecność bardzo młodych i bardzo starych pasażerów może mieć znaczenie przy analizie przeżycia.

2. **Liczba rodzeństwa/małżonków na pokładzie (SibSp)**

- Mediana i pierwszy kwartyl to 0, co oznacza, że **ponad połowa pasażerów podróżowała bez rodzeństwa lub współmałżonka**.
- Maksymalna wartość to 8, co wskazuje na kilka dużych rodzin na pokładzie.

3.**Liczba rodziców/dzieci na pokładzie (Parch)**

- Tutaj także mediana to 0, czyli **większość pasażerów nie podróżowała z rodzicem lub dzieckiem**.
- Maksymalna wartość wynosi 6, co podobnie jak przy SibSp, wskazuje na obecność dużych rodzin.

4.**Rozmiar rodziny (FamilySize)**

- Zmienna ta łączy SibSp i Parch, doliczając samego pasażera.
- **Większość pasażerów podróżowała samotnie lub z małą rodziną** – mediana to 1, średnia to ok. 1.9.
- Największa rodzina miała **11 członków** na pokładzie.

- Może to być istotne dla analizy przeżycia – osoby podróżujące z rodziną mogły mieć większe wsparcie.

W celu oceny kształtu rozkładu analizowanych zmiennych obliczono również współczynniki skośności i kurtozy:

1. Age (Wiek):

- **Skośność: 0.43** (lekka skośność w prawo).
- **Kurtoza: 0.95** (rozklad zbliżony do normalnego, bliski rozkładowi normalnemu).
- Rozkład wieku pasażerów jest stosunkowo zbliżony do rozkładu normalnego. Wartość skośności wskazuje na niewielką tendencję do młodszych pasażerów (choć efekt jest mały). Kurtoza bliska 3 sugeruje, że rozkład nie odbiega znacząco od normalności.

2. Fare (Opłata):

- **Skośność: 4.77** (silna skośność w prawo).
- **Kurtoza: 33.12** (bardzo wysoka kurtoza).
- Rozkład opłat jest silnie skośny w prawo, co oznacza, że większość pasażerów zapłaciła niższe ceny, a tylko kilku zapłaciło bardzo wysokie kwoty. Bardzo wysoka kurtoza wskazuje na obecność wartości skrajnych (outliers), co sugeruje, że niektóre osoby zapłaciły znacznie więcej niż inni pasażerowie.

3. SibSp (Liczba rodzeństwa / małżonków na pokładzie):

- **Skośność: 3.68** (bardzo silna skośność w prawo).
- **Kurtoza: 17.73** (bardzo wysoka kurtoza).
- Większość pasażerów nie miała rodzeństwa ani małżonka na pokładzie. Skośność wskazuje na to, że istnieje niewielka liczba pasażerów z wieloma krewnymi na pokładzie, a wysoka kurtoza sugeruje obecność kilku osób, które podróżowały z wieloma członkami rodziny.

4. Parch (Liczba rodziców / dzieci na pokładzie):

- **Skośność: 2.74** (silna skośność w prawo).
- **Kurtoza: 9.69** (wysoka kurtoza).
- Podobnie jak w przypadku "SibSp", większość pasażerów nie miała dzieci ani rodziców na pokładzie. Skośność wskazuje na obecność kilku pasażerów z większą liczbą dzieci lub rodziców. Wysoka kurtoza wskazuje na wartości skrajne.

5. FamilySize (Wielkość rodziny, czyli SibSp + Parch):

- **Skośność: 2.72** (silna skośność w prawo).

- **Kurtoza: 9.07** (wysoka kurtoza).
- Większość pasażerów podróżowała samotnie lub z jedną osobą. Skośność wskazuje na to, że część pasażerów podróżowała z większymi rodzinami. Kurtoza wskazuje na obecność kilku większych rodzin w zbiorze danych.

2.Rozkład zmiennych kategorycznych

1. Survived (Przeżycie)

- **550** pasażerów nie przeżyło katastrofy (**0**).
- **341** pasażerów przeżyło katastrofę (**1**).
- Większość pasażerów **nie przeżyła** katastrofy, co jest zgodne z historycznymi danymi o Titanicu.

2. Pclass (Klasa biletu)

- **216** pasażerów podróżowało w **klasie 1**.
- **184** pasażerów w **klasie 2**.
- **491** pasażerów w **klasie 3**.
- Najwięcej pasażerów podróżowało w **klasie 3**, co może wskazywać na dużą liczbę osób z niższych klas społecznych.

3. Sex (Płeć)

- **314** pasażerów to **kobiety** (**female**).
- **577** pasażerów to **mężczyźni** (**male**).
- Większość pasażerów stanowili **mężczyźni**, co odzwierciedla faktyczny rozkład pasażerów.

4. Embarked (Port zaokrętowania)

- **168** pasażerów wsiadło na pokład w **Cherbourg** (**C**).
- **77** pasażerów wsiadło w **Queenstown** (**Q**).
- **646** pasażerów wsiadło w **Southampton** (**S**).
- Większość pasażerów wsiadła w **Southampton**.

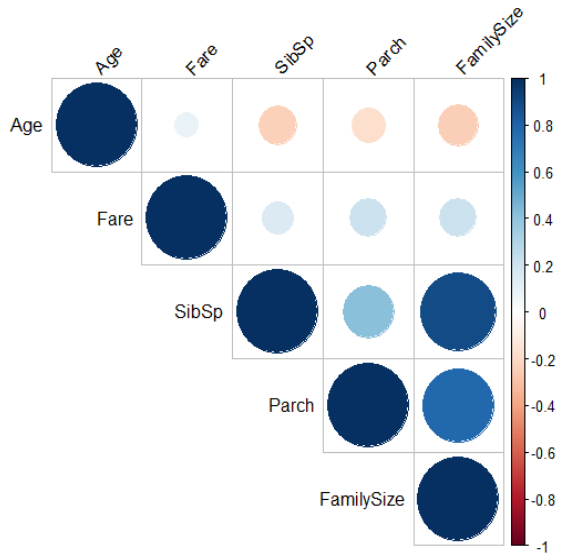
3.Analiza korelacji

Korelacja zmiennych numerycznych

W celu zbadania zależności pomiędzy zmiennymi liczbowymi, utworzono macierz korelacji oraz odpowiadający jej wykres przy użyciu pakietu corrplot. Analizowane zmienne to: Age, Fare, SibSp, Parch oraz FamilySize.

	Age	Fare	SibSp	Parch	FamilySize
Age	1.00000000	0.09156609	-0.2326246	-0.1791909	-0.2485117
Fare	0.09156609	1.00000000	0.1596510	0.2162249	0.2171384
SibSp	-0.2326246	0.1596510	1.0000000	0.4148377	0.8907117
Parch	-0.1791909	0.2162249	0.4148377	1.0000000	0.7831108
FamilySize	-0.2485117	0.2171384	0.8907117	0.7831108	1.0000000

- Istnieje wyraźna korelacja między zmiennymi dotyczącymi liczby członków rodziny, zwłaszcza **SibSp** i **Parch**, które razem tworzą zmienną **FamilySize**. Wartości korelacji wskazują, że pasażerowie, którzy podróżują w większych rodzinach, mają wyższą liczbę rodzeństwa i małżonków oraz rodziców i dzieci.
- Zmienna **Fare** (cena biletu) wykazuje pozytywne korelacje z **SibSp**, **Parch** i **FamilySize**, co sugeruje, że pasażerowie, którzy płacą wyższe ceny za bilety, mogą podróżować w towarzystwie większej liczby osób.
- **Age** (wiek) wykazuje negatywne korelacje z liczbą członków rodziny, co może oznaczać, że młodsze osoby częściej podróżują w większym gronie rodzinnym, podczas gdy starsi pasażerowie podróżują samotnie.



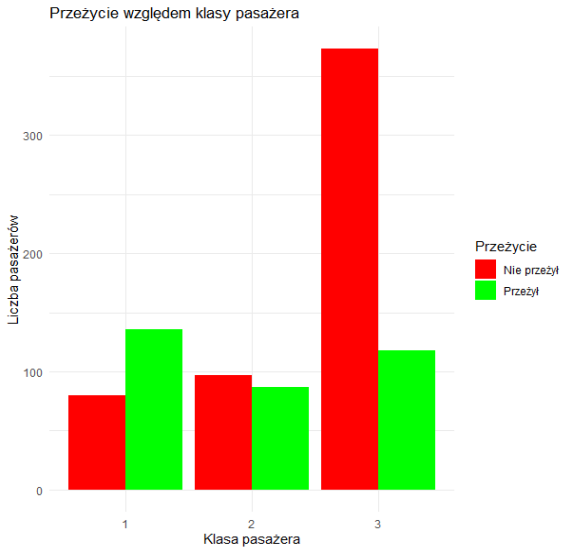
Korelacja zmiennych kategorycznych

Z kolei zmienne kategoryczne, takie jak Survived, Pclass, Sex, wymagają innej metody analizy korelacji. W tym przypadku zastosujemy test chi-kwadrat, który pozwala na ocenę, czy istnieje zależność pomiędzy zmiennymi kategorycznymi, np. czy klasa(Pclass) ma wpływ na przeżycie (Survived).

Wynik testu chi-kwadrat dla Pclass:

- **X-squared = 104.2**: Jest to wartość statystyki chi-kwadrat, która mierzy, jak bardzo różnią się obserwowane wartości od wartości oczekiwanych.
- **df = 2**: To liczba stopni swobody, która zależy od liczby kategorii w badanych zmiennych. W tym przypadku mamy 3 kategorie w zmiennej Pclass i 2 kategorie w zmiennej Survived, więc stopni swobody wynosi (3-1) * (2-1) = 2.
- **p-value < 2.2e-16**: P-wartość jest mniejsza niż 2.2e-16, co oznacza, że wynik jest bardzo statystycznie istotny. Możemy więc odrzucić hipotezę zerową, że nie ma zależności między zmiennymi Survived i Pclass.

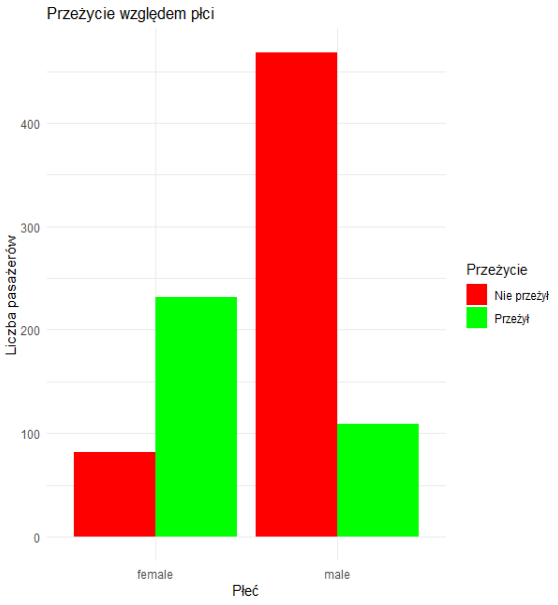
Na podstawie testu chi-kwadrat możemy stwierdzić, że istnieje statystycznie istotna zależność między klasą biletu (Pclass) a tym, czy pasażer przeżył (Survived). Wynik ten jest bardzo istotny, ponieważ p-wartość jest znacznie mniejsza od standardowego poziomu istotności (0.05). W kontekście analizy danych dotyczących Titanica, może to sugerować, że pasażerowie z wyższych klas (np. 1. klasa) mieli większe szanse na przeżycie niż pasażerowie z niższych klas, co możemy zobaczyć na wykresie przeżycia względem klasy.



Wynik testu chi-kwadrat dla zmiennych **Sex** i **Survived** jest następujący:

- Statystyka chi-kwadrat: **269.52**
- Stopnie swobody: df=**28**
- p -value < **2.2e-16** (bardzo mała wartość, znacznie mniejsza niż 0.05)

Test wykazuje **istotną statystycznie zależność** między płcią a przeżyciem. Oznacza to, że płeć ma duży wpływ na szanse przeżycia pasażerów Titanica.



Podsumowanie

W przeprowadzonej analizie danych pasażerów Titanica zidentyfikowano istotne czynniki, które mogły wpływać na przeżycie podczas katastrofy. Proces preprocessingu pozwolił na oczyszczenie i przygotowanie danych do analizy, m.in. poprzez usunięcie nieistotnych kolumn, uzupełnienie brakujących danych oraz transformację zmiennych kategorycznych.

Analiza statystyczna wykazała, że:

- **Płeć** miała istotny wpływ na przeżycie – kobiety przeżywały znacznie częściej niż mężczyźni.
- **Klasa biletu (Pclass)** również silnie korelowała z przeżyciem – pasażerowie z 1. klasy mieli największe szanse na ocalenie.
- **Rozmiar rodziny (FamilySize)** wpływał na przeżycie – osoby podróżujące samotnie miały niższe szanse niż te z małą rodziną.
- **Wiek** i **cena biletu (Fare)** wykazywały umiarkowane korelacje z innymi zmiennymi, ale ich wpływ na przeżycie wymagałby dalszej analizy.

Wyniki testu chi-kwadrat potwierdziły statystycznie istotne zależności między zmiennymi kategorycznymi, w szczególności pomiędzy **klasą biletu a przeżyciem**. Dodatkowo, wizualizacja macierzy korelacji pozwoliła lepiej zrozumieć zależności pomiędzy zmiennymi liczbowymi.

Na podstawie danych z pliku **titanic_new.csv** można przeprowadzić szereg analiz eksploracyjnych, statystycznych oraz predykcyjnych. Przykładowe kierunki analizy to:

1. **Analiza przeżycia (Survival Analysis)**
   Celem jest identyfikacja czynników wpływających na przeżycie pasażerów. Kluczowe zmienne: Survived, Sex, Pclass, Age, Fare, Embarked, FamilySize. Analizę można wspomóc testami chi-kwadrat oraz badaniem korelacji.
2. **Analiza demograficzna**
   Polega na poznaniu struktury demograficznej pasażerów według wieku, płci, klasy podróży czy portu zaokrętowania. Wykorzystuje się zmienne takie jak Age, Sex, Pclass, Embarked, FamilySize.
3. **Analiza zależności między zmiennymi**
   Można badać związki pomiędzy zmiennymi liczbowymi, np. czy większa liczba członków rodziny wpływała na wysokość zapłaconego biletu (Fare). W tym celu można wykorzystać macierz korelacji oraz wizualizacje (np. wykresy korelacyjne).

4. **Modelowanie predykcyjne (np. klasyfikacja)**
   Na podstawie dostępnych cech pasażera można zbudować model przewidujący prawdopodobieństwo przeżycia. W analizie mogą zostać użyte takie zmienne jak: Sex, Pclass, Age, Fare, Embarked, FamilySize.

5. **Testy statystyczne**
   Przykładowo można sprawdzić, czy klasa podróży (Pclass) lub płeć (Sex) miały istotny wpływ na przeżycie. Do tego celu stosuje się m.in. test chi-kwadrat dla zmiennych kategorycznych.

Bibliografia:

https://statystyka.online/teoria/

Kod:

```
# Wczytanie danych

dane <- read.csv("C:\\Users\\Dell\\Desktop\\Semestr 2\\R\\9_Preprocessing\\titanic_new.csv", stringsAsFactors = FALSE)

View(dane)


# Sprawdzenie struktury danych

str(dane)


# Konwersja wieku na typ numeryczny

dane$Age <- as.numeric(dane$Age)


# Usunięcie zbędnych kolumn

dane <- dane[ , !(names(dane) %in% c("PassengerId", "Passenger.Id", "Name", "Ticket", "Cabin"))]


# Konwersja wybranych kolumn do typu factor

dane$Survived <- as.factor(dane$Survived)

dane$Pclass   <- as.factor(dane$Pclass)

dane$Sex      <- as.factor(dane$Sex)

dane$Embarked <- as.factor(dane$Embarked)


# Uzupełnienie brakujących wartości w kolumnie Age średnią

dane$Age[is.na(dane$Age)] <- mean(dane$Age, na.rm = TRUE)


# Uzupełnienie braków w Embarked najczęściej występującą wartością

najczestszy_port <- names(sort(table(dane$Embarked), decreasing = TRUE))[1]
```

```
dane$Embarked[is.na(dane$Embarked) | dane$Embarked == ""] <-
najczestszy_port

# Dodanie kolumny FamilySize (łącznie z pasażerem)
dane$FamilySize <- dane$SibSp + dane$Parch + 1

# Statystyki opisowe dla zmiennych liczbowych
summary(dane[, sapply(dane, is.numeric)])

# Czyszczenie danych - standaryzacja błędnych wartości
dane$Sex <- tolower(dane$Sex)
dane$Sex[dane$Sex %in% c("female", "feemale", "f2emale")] <-
"female"
dane$Sex[dane$Sex %in% c("male", "mal3e", "mal4e", "malle",
"malwe")] <- "male"
dane$Sex <- as.factor(dane$Sex)

dane$Embarked <- toupper(dane$Embarked)
dane$Embarked <- as.factor(dane$Embarked)

# Sprawdzenie rozkładu zmiennych
table(dane$Sex)
table(dane$Pclass)
table(dane$Embarked)

# Poprawienie wartości w kolumnie Survived (musi być 0 lub 1)
dane$Survived <- as.numeric(as.character(dane$Survived))
dane$Survived[dane$Survived > 1] <- 1
dane$Survived[dane$Survived < 0] <- 0
dane$Survived <- as.factor(dane$Survived)
table(dane$Survived)
```

```
# Obliczenie macierzy korelacji dla wybranych zmiennych
liczbowych

correlation_matrix <- cor(dane[c("Age", "Fare", "SibSp",
"Parch", "FamilySize")], use = "complete.obs")


# Sprawdzenie typu danych przed wykresem

str(dane)


# Usunięcie znaków niebędących liczbami z Fare, jeśli
występują

dane$Fare <- as.numeric(gsub("'", "", dane$Fare))


# Wykres macierzy korelacji

library(corrplot)

corrplot(correlation_matrix, method = "circle", type =
"upper", tl.col = "black", tl.srt = 45)


# Test chi-kwadrat: zależność między przetrwaniem a klasą

chisq_test_2 <- chisq.test(table(dane$Survived, dane$Pclass))

print(chisq_test_2)

# Test chi-kwadrat między płcią a przeżyciem

chisq_test_sex <- chisq.test(table(dane$Sex, dane$Survived))

print(chisq_test_sex)


# Wczytanie pakietu e1071 do analizy skośności i kurtozy

library(e1071)


# Obliczenie skośności i kurtozy dla zmiennych liczbowych

skewness_age <- skewness(dane$Age, na.rm = TRUE)

kurtosis_age <- kurtosis(dane$Age, na.rm = TRUE)
```

```
skewness_fare <- skewness(dane$Fare, na.rm = TRUE)
kurtosis_fare <- kurtosis(dane$Fare, na.rm = TRUE)

skewness_sibsp <- skewness(dane$SibSp, na.rm = TRUE)
kurtosis_sibsp <- kurtosis(dane$SibSp, na.rm = TRUE)

skewness_parch <- skewness(dane$Parch, na.rm = TRUE)
kurtosis_parch <- kurtosis(dane$Parch, na.rm = TRUE)

skewness_familysize <- skewness(dane$FamilySize, na.rm = TRUE)
kurtosis_familysize <- kurtosis(dane$FamilySize, na.rm = TRUE)

# Wyświetlenie wyników
cat("Skośność (Age):", skewness_age, "Kurtoza (Age):",
kurtosis_age, "\n")
cat("Skośność (Fare):", skewness_fare, "Kurtoza (Fare):",
kurtosis_fare, "\n")
cat("Skośność (SibSp):", skewness_sibsp, "Kurtoza (SibSp):",
kurtosis_sibsp, "\n")
cat("Skośność (Parch):", skewness_parch, "Kurtoza (Parch):",
kurtosis_parch, "\n")
cat("Skośność (FamilySize):", skewness_familysize, "Kurtoza
(FamilySize):", kurtosis_familysize, "\n")
#wykresy
library(ggplot2)

ggplot(dane, aes(x = Sex, fill = Survived)) +
  geom_bar(position = "dodge") +
  labs(title = "Przeżycie względem płci",
       x = "Płeć",
```

```
       y = "Liczba pasażerów",
       fill = "Przeżycie") +
  scale_fill_manual(values = c("red", "green"),
                    labels = c("Nie przeżył", "Przeżył")) +
  theme_minimal()

ggplot(dane, aes(x = Pclass, fill = Survived)) +
  geom_bar(position = "dodge") +
  labs(title = "Przeżycie względem klasy pasażera",
       x = "Klasa pasażera",
       y = "Liczba pasażerów",
       fill = "Przeżycie") +
  scale_fill_manual(values = c("red", "green"),
                    labels = c("Nie przeżył", "Przeżył")) +
  theme_minimal()
```

Król Martyna