Analiza danych pasażerów Titanica

Marta Rauch

8 czerwca 2023

Spis treści

1 Wprowadzenie

Tematem naszej pracy jest analiza danych na podstawie pasażerów Titanica. Dane zostały pobrane z pakietu wbudowanego w środowisko programistyczne R. Zaczynamy od wczytania danych "titanic_train" z pakietu Titanic, które zostały stworzone w celach edukacyjnych. Przed rozpoczęciem analizy, przyjrzyjmy się naszemu zbiorowi danych.

Wyświetlmy zatem nasz zbiór danych:

Description: df [891 × 12]

PassengerId <int>	Survived <int>	Pclass <int>	Name <chr>	Sex <chr>	Age <dbl>	SibSp <int>	Parch <int>	Ticket <chr>
1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599
3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803
5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450
6	0	3	Moran, Mr. James	male	*NA*	0	0	330877
7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463
8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736

1–10 of 891 rows | 1–9 of 12 columns

Previous 1 2 3 4 5 6 ... 90 Next

Description: df [891 × 12]

Name <chr>	Sex <chr>	Age <dbl>	SibSp <int>	Parch <int>	Ticket <chr>	Fare <dbl>	Cabin <chr>	Embarked <chr>
Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500		S
Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250		S
Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S
Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500		S
Moran, Mr. James	male	*NA*	0	0	330877	8.4583		Q
McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46	S
Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750		S
Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333		S
Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0708		C

1–10 of 891 rows | 4–12 of 12 columns

Previous 1 2 3 4 5 6 ... 90 Next

Widać, że analizowane przez nas dane przedstawione są w formie tabeli o 891 wierszach i 12 kolumnach. Liczba wierszy odpowiada liczbie pasażerów, których dane zostały zebrane.

Warto zauważyć, że oryginalnie na statku Titanic podróżowało około 1500 pasażerów, natomiast w tym konkretnym zbiorze nie wszystkie dane zostały uwzględnione. Może to wynikać z różnych czynników, jak na przykład brakujące czy utracone informacje po katastrofie statku.

2 Dane

Przyjrzyjmy się analizowanemu przez nas zbiorowi danych. Wiemy już, że liczba wierszy odpowiada liczbie pasażerów. Pozostało więc sprawdzić, co określają poszczególne kolumny tabeli.

Zbiór danych Titanic składa się z następujących kolumn:

1. **PassengerId** - jest to liczba porządkowa, określająca numer pasażera,
2. **Survived** - określa, czy dany pasażer przeżył (wartość 1), czy zginął (wartość 0); zatem w tej kolumnie stosowany jest zapis binarny.
3. **Pclass** - oznacza klasę podróżną pasażera; 1 to klasa wyższa, 2 średnia, a 3 niższa.
4. **Name** - zawiera imię i nazwisko pasażera.
5. **Sex** - określa płeć pasażera.
6. **Age** - dostarcza informacji na temat wieku pasażera.
7. **SibSp** - oznacza liczbę rodzeństwa lub małżonków podróżujących razem z pasażerem.
8. **Parch** - wskazuje liczbę rodziców lub dzieci podróżujących razem z pasażerem.
9. **Ticket** - zawiera numer danego biletu.
10. **Fare** - dostarcza informacji na temat ceny biletu.
11. **Cabin** - zawiera informacje na temat numeru kabiny, w której przebywał pasażer (większość wartości w tej kolumnie jest nieznana).
12. **Embarked** - wskazuje port, z którego dany pasażer wsiadał na statek.

3 Wstępna obróbka danych

Kolumna Age posiada wiele wartości nieokreślonych (NA). W celu zachowania ogólnego charakteru danych i zminimalizowania wpływu brakujących wartości na analizę, zastąpimy je średnią wartością wieku pasażerów Titanica. Takie podejście nie wpłynie na pogorszenie jakości przeprowadzanej przez nas analizy, gdyż wiek jest wartością losową.

4 Statystyki opisowe

W tej części przejdziemy do analizy danych na podstawie statystyk opisowych, takich jak: średnia, odchylenie standardowe, mediana, IQR, zakres (min-max). Taką analizę przeprowadzimy dla każdej ze zmiennych. Kolumny takie jak PassengerId, Name, Ticket oraz Cabin nie dostarczają nam żadnych danych ilościowych, stąd nie bierzemy ich pod uwagę. Przeanalizujemy pozostałe zmienne, aby uzyskać pełniejsze zrozumienie danych i wyciągnąć przydatne wnioski.

Danych dotyczących średniej, mediany, IQR (rozstępu międzykwartylowego), mediany oraz zakresu (minimum oraz maksimum) dostarczy nam funkcja *summary()* wbudowana w środowisko obliczeniowe R. Odchylenie standardowe natomiast obliczymy za pomocą funkcji *var()* (wariancja), którą spierwiastkujemy.

4.1 Survived - ocalali

Tak prezentują się wyniki zastosowania funkcji *summary()* na danych z kolumny Survived:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3838  1.0000  1.0000
```

4.1.1 Średnia

Średnia wartość w tej kolumnie wyniosła 0.3838. Wiemy, że dane pochodzące z kolumny survived są binarne. Stąd możemy stwierdzić, że średnio 38, 38% pasażerów zostało uratowanych. W naszym przypadku oznacza to, że ocalono 341 osób, natomiast nie udał ocalić się 550 pasażerów.

4.1.2 Odchylenie standardowe

````
```{r}
var(titanic$Survived)^(1/2)
```
````

```
[1] 0.4865925
```

Odchylenie standardowe danych w tej kolumnie to w przybliżeniu 0.49. Jest to wartość bliska 0.5, co sugeruje, że wartości są równomiernie rozłożone wokół średniej. W naszym przypadku oznacza to, że blisko połowa pasażerów ocalała.

4.1.3 Mediana

Mediana wynosi 0, co wskazuje na to, że większość pasażerów nie została ocalona.

4.1.4 IQR

IQR to rozstęp międzykwartylowy, czyli różnica między trzecim a pierwszym kwartylem. Ponieważ pomiędzy tymi kwartylami znajduje się z definicji 50% wszystkich obserwacji (położonych centralnie w rozkładzie), dlatego im większa szerokość rozstępu ćwiartkowego, tym większe zróżnicowanie cechy.

W przypadku kolumny Survived, która oznacza zmienną binarną, wartość IQR równa 1 sugeruje, że 50% obserwacji ma wartość 1 (ocalali pasażerowie), a pozostałe 50% z wartością 0 nie zostało ocalonych.

4.1.5 Zakres

Zmienna Survived przyjmuje tylko wartości 0 lub 1, stąd minimum wynosi 0, a maksimum 1.

4.2 Pclass - klasa podróżna

W przypadku zmiennej Pclass, statystyki opisowe prezentują się następująco:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.000   2.000   3.000   2.309   3.000   3.000
```

4.2.1 Średnia

W przypadku zmiennej Pclass, średnia wartość wynosi 2.309. Sugeruje nam to, że średnio pasażerowie wybierali klasę podróżną średnią (2). Niekoniecznie jednak odzwierciedla nam to liczebność poszczególnych klas.

4.2.2 Odchylenie standardowe

````
```{r}
var(titanic$Pclass)^(1/2)
```
````

```
[1] 0.8360712
```

Odchylenie standardowe przyjmuje wartość równą 0.84. Wskazuje to na rozproszenie klas podróży wokół średniej. W takim razie pasażerowie podróżowali zarówno w niższych, jak i wyższych klasach.

4.2.3 Mediana

Mediana przyjmuje wartość 3, zatem przynajmniej połowa osób podróżowała w 3, czyli niższej klasie.

4.2.4 IQR

Rozstęp międzykwartylowy wynosi 1. Oznacza to, że 50% wartości klas podróży koncentruje się w przedziale o szerokości 1. Natomiast rozstęp międzykwartylowy nie dostarcza nam bezpośredniej informacji o liczebności poszczególnych klas, a jedynie o rozstępie między trzecim a pierwszym kwartylem.

4.2.5 Zakres

Minimum dla tego zbioru wynosi 1, natomiast maksimum 3. Zatem na statku były 3 klasy. Obliczyliśmy liczbę pasażerów w poszczególnych klasach:

```{r}
sum(titanic$Pclass==1)
sum(titanic$Pclass==2)
sum(titanic$Pclass==3)
```

```
[1] 216
[1] 184
[1] 491
```

Wiemy już zatem, że liczba pasażerów w pierwszej klasie wyniosła 216; w drugiej klasie podróżowały 184 osoby, natomiast większość pasażerów podróżowała w klasie 3 - było to 491 osób.

4.3 Sex - płeć

Z danych zawartych w kolumnie płeć otrzymaliśmy informację, że liczba kobiet na statku wynosiła 314, natomiast mężczyzn było 577. Zatem zdecydowana większość pasażerów to mężczyźni.

```{r}
sum(titanic$Sex=='female')
sum(titanic$Sex=='male')
```

```
[1] 314
[1] 577
```

4.4 Age - wiek

Statystyki opisowe dla zmiennej Age prezentują się następująco:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.42   20.12   28.00   29.70   38.00   80.00
```

4.4.1 Średnia

Średnia wieku pasażerów statku wynosi 29.7 lat.

4.4.2 Odchylenie standardowe

```{r}
var(titanic$Age)^(1/2)
```

```
[1] 13.00202
```

Wartość odchylenia standardowego wynosi w przybliżeniu 13, co wskazuje na dosyć spore rozproszenie wieku wokół średniej.

4.4.3 Mediana

Mediana wieku to 28 lat, zatem 50% pasażerów miało mniej niż 28 lat, a drugie 50% - więcej.

4.4.4 IQR

Rozstęp międzykwartylowy osiąga wartość 17.85, co wskazuje na rozpiętość wieku w analizowanej grupie danych. Oznacza to, że 50% osób w tej grupie ma wiek mieszczący się w zakresie od pierwszego kwartyla (20.12) do trzeciego kwartyla (38). Jeśli średnia wieku wynosi 29.7 lat, IQR nie odzwierciedla bezpośrednio średniej wieku, ale informuje nas o rozproszeniu wieku wśród analizowanej grupy danych.

4.4.5 Zakres

Minimum w tym zbiorze to 0.42, natomiast maksimum 80. Wnioskujemy zatem, że najmłodszy pasażer na statku miał około 5 miesięcy, natomiast najstarszy aż 80 lat.

4.5 SibSp - liczba rodzeństwa/małżonków podróżujących razem z pasażerem

Sprawdzamy statystyki opisowe dla kolumny SibSp:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   0.000   0.000   0.523   1.000   8.000
```

4.5.1 Średnia

Wartość średnia w kolumnie SibSp wynosi 0.52. Oznacza to, że pasażerowie Titanica mieli średnio 0.52 rodzeństwa lub małżonków podróżujących razem z nimi. Wartość nie jest bliska jedynce, zatem możemy wnioskować, że większość pasażerów nie podróżowała z małżonkami czy rodzeństwem.

4.5.2 Odchylenie standardowe

````
```{r}
var(titanic$SibSp)^(1/2)
```
````

```
[1] 1.102743
```

Wartość odchylenia standardowego wynosi 1.10, co wskazuje na dosyć spore zróżnicowanie liczby rodzeństwa lub małżonków podróżujących z pasażerami statku.

4.5.3 Mediana

Mediana jest równa zero, zatem ponad połowa pasażerów nie miała ze sobą na statku rodzeństwa czy małżonków.

4.5.4 IQR

Rozstęp międzykwartylowy w tym przypadku osiąga wartość 1, zatem wartości koncertują się głównie w przedziale od zera do jedynki.

4.5.5 Zakres

Najmniejsza osiągalna wartość w tym zbiorze to 0, natomiast największa to 8. Jak wcześniej stwierdziliśmy, większość pasażerów podróżowała bez rodzeństwa lub małżonków, natomiast niektórzy z pasażerów mieli ich ze sobą na statku aż 8.

4.6 Parch - liczba rodziców/dzieci podróżujących razem z pasażerem

Tak prezentują się statystyki opisowe dla kolumny Parch:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.3816  0.0000  6.0000 
```

4.6.1 Średnia

Średnia wartość dla zmiennej Parch wynosi 0.38, co wskazuje, że pasażerowie w tym zbiorze danych mieli średnio mniej niż 1 rodzica lub dziecka podróżującego z nim. Sugeruje nam to, że prawdopodobnie większość pasażerów nie miała ze sobą żadnego rodzica/dzieci lub miała ich ze sobą bardzo małą ilość na pokładzie.

4.6.2 Odchylenie standardowe

````
```{r}
var(titanic$Parch)^(1/2)
```
````

```
[1] 0.8060572
```

Wartość odchylenia standardowego, równa w przybliżeniu 0.81 sugeruje, że wartości w tej kolumnie mają stosunkowo dużą zmienność wokół średniej wartości.

4.6.3 Mediana

Mediana wynosi 0, zatem faktycznie przynajmniej połowa pasażerów nie miała ze sobą żadnego rodzica/dziecka podróżującego wraz z nimi na pokładzie.

4.6.4 IQR

Wartość rozstępu międzykwartylowego w tym przypadku wynosi 0. Oznacza to, że środkowe 50% obserwacji w tej grupie ma tę samą wartość dla badanej cechy. Można to interpretować jako bardzo małe zróżnicowanie w tej konkretnej kolumnie, ponieważ większość osoby w tej grupie nie miała ze sobą na pokładzie żadnego rodzica/dzieci.

4.6.5 Zakres

Zakres wartości w tej kolumnie wynosi od 0 do 6. Zatem w analizowanym zbiorze pasażerowie mieli najmniej zero rodziców/dzieci podróżujących z nim na pokładzie, a najwięcej - sześcioro.

4.7 Fare - opłata za bilet

Prezentujemy statystyki opisowe w kolumnie Fare:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00    7.91   14.45   32.20   31.00  512.33
```

4.7.1 Średnia

Średnia wartość dla tej cechy wynosi 32.20, co wskazuje, że pasażerowie średnio za bilet płacili 32.20 jednostek (prawdopodobnie były to dolary amerykańskie).

4.7.2 Odchylenie standardowe

```
```{r}
var(titanic$Fare)^(1/2)
```
```

```
[1] 49.69343
```

Odchylenie standardowe dla tej zmiennej wynosi 49.69, co sugeruje, że istnieje duże zróżnicowanie w kwotach, które pasażerowie płacili za bilety.

4.7.3 Mediana

Mediana w zbiorze wynosi 14.45, zatem 50% pasażerów za bilet zapłaciło mniej niż 14.45$, natomiast drugie 50% - zapłaciło więcej.

4.7.4 IQR

Zakres międzykwartylowy przyjmuje wartość 23.09, co wykazuje, że środkowe 50% pasażerów za bilet zapłaciło kwotę mieszczącą się w przedziale o szerokości 23.09.

4.7.5 Zakres

Zakres w tym przypadku wynosi od 0 aż do 512.33. Oznacza to, że niektórzy pasażerowie dostali bilet za darmo, natomiast niektórzy musieli zapłacić za niego aż 512.33$.

4.8 Embarked - port, z którego wsiadał pasażer

Z kolumny Embarked odczytujemy dane na temat portów, z których wsiadali pasażerowie. Na pokład można było wsiąść z trzech portów: Cherbourg, Queenstown oraz Southampton. Tak przedstawiają się statystyki:

- z portu w Cherbourg wsiadło 168 osób,
- na pokład w Queenstown wsiadło 77 pasażerów,
- na statek w Southampton weszły 644 osoby.

5 Wizualizacja danych

Tworzenie wizualizacji będzie kluczowe dla lepszego zrozumienia danych dotyczących wieku, płci, klasy podróżnej oraz zależności między wiekiem a ceną biletu dla ocalałych pasażerów na Titanicu. Poprzez wizualizacje będziemy w stanie przekazać wyniki w sposób przystępny i zrozumiały. Wykorzystanie wizualnych prezentacji pozwoli nam łatwiej dostrzec wzorce, trendy oraz potencjalne odstępstwa, co z kolei przyczyni się do bardziej dogłębnej analizy danych dotyczących pasażerów Titanica.

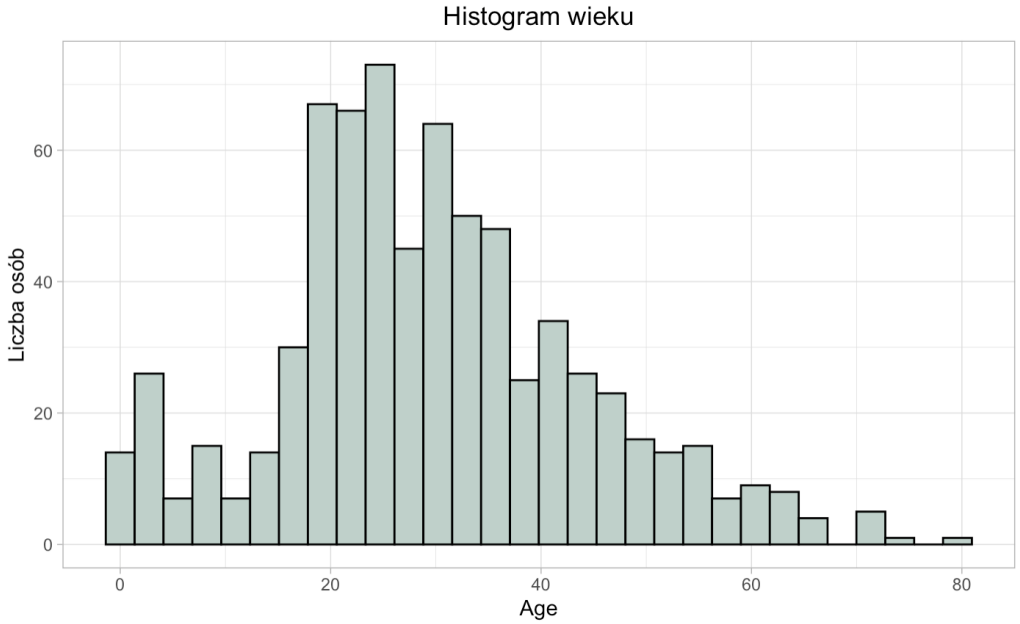
Tworzenie wizualizacji umożliwi nam zobaczenie i zrozumienie różnych aspektów w bardziej przystępny sposób. Na przykład, poprzez odpowiednie wykresy czy wykresy słupkowe, będziemy mogli porównać wiek, płeć i klasę podróżną w jednym miejscu, co pozwoli nam wyciągnąć wnioski na temat demografii na Titanicu. Ponadto, wizualizacje zależności między wiekiem a ceną biletu dla ocalałych pasażerów pozwolą nam zbadać ewentualne trendy lub korelacje.

Korzystanie z wizualizacji w analizie danych związanych z pasażerami Titanica pozwoli nam spojrzeć na te informacje w sposób bardziej kompleksowy i umożliwi nam odkrycie istotnych wzorców, które mogą pomóc w dalszych badaniach.

5.1 Wiek pasażerów

Wizualizacja wieku pasażerów pozwoli nam lepiej zrozumieć demografię na pokładzie Titanica. Wykorzystując dostępne dane dotyczące wieku, możemy stworzyć histogram, który przedstawi rozkład wieku wśród badanych osób. Histogram to wykres, który ilustruje częstość występowania różnych wartości w określonych przedziałach wiekowych.

Analizując histogram, będziemy w stanie dostrzec wzorce, trendy i potencjalne odstępstwa związane z wiekiem pasażerów. Ta wizualizacja pomoże nam lepiej zrozumieć strukturę wiekową grupy badanych osób na pokładzie Titanica.

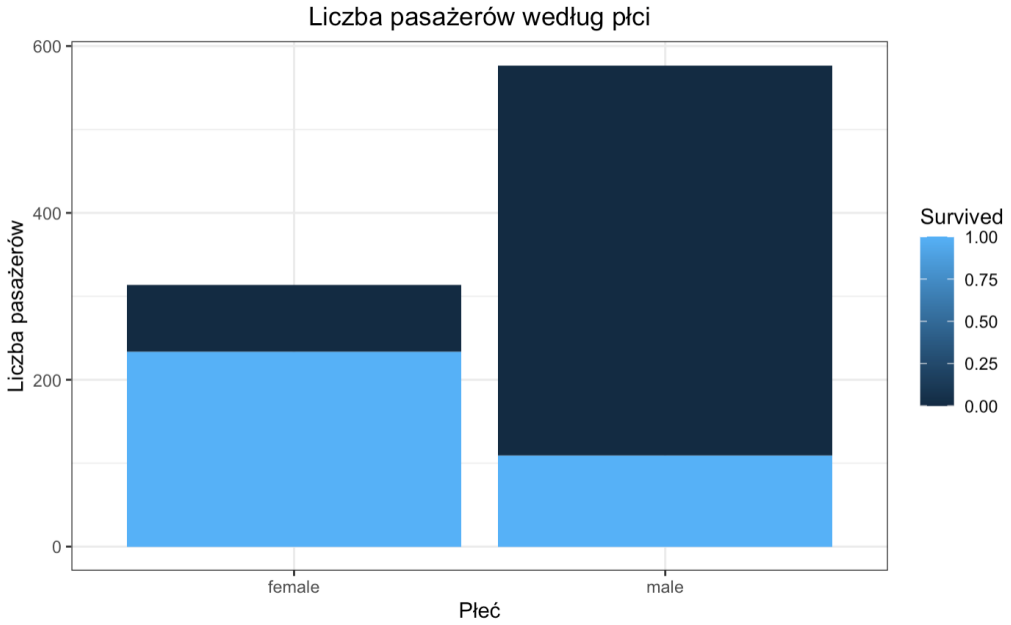


Histogram przedstawiający wiek pasażerów Titanica został podzielony na przedziały wiekowe co 10 lat, z zakresem wieku od 0 do 80 lat na osi poziomej (osi X). Analizując ten histogram, możemy zauważyć, że największa liczba pasażerów skupiała się w przedziale wiekowym między 20 a 30 lat. Oznacza to, że grupa wiekowa między 20 a 30 rokiem życia była najliczniejsza wśród pasażerów Titanica. Z drugiej strony, najmniejsza liczba pasażerów występowała w przedziale wiekowym od 70 do 80 lat. Sugeruje to, że osoby w podeszłym wieku były mniej licznie reprezentowane na statku.

Ponadto, histogram przedstawiający wiek pasażerów Titanica jest prawostronnie asymetryczny. Oznacza to, że rozkład wieku ma wyższe wartości skupione w lewej części histogramu (młodszy wiek) i stopniowo malejące wartości w prawej części (starszy wiek). Największa liczba pasażerów koncentruje się w młodszych grupach wiekowych, co jest widoczne na wykresie poprzez wyższe słupki w przedziałach wiekowych od 20 do 30 lat. Wraz z postępującym wiekiem, liczba pasażerów maleje, co jest zilustrowane niższymi słupkami w przedziałach wiekowych od 70 do 80 lat. Takie asymetryczne rozłożenie sugeruje, że na Titanicu była większa liczba młodszych pasażerów w porównaniu do osób starszych.

5.2 Płeć pasażerów

W tej części przeanalizujemy przy pomocy wykresu słupkowego rozkład płci pasażerów Titanica. Pozwoli nam to spostrzec, jak różne grupy płciowe były reprezentowane na statku. Ponadto sprawdzimy, jaka część kobiet i mężczyzn ocalała, a jaka zginęła.



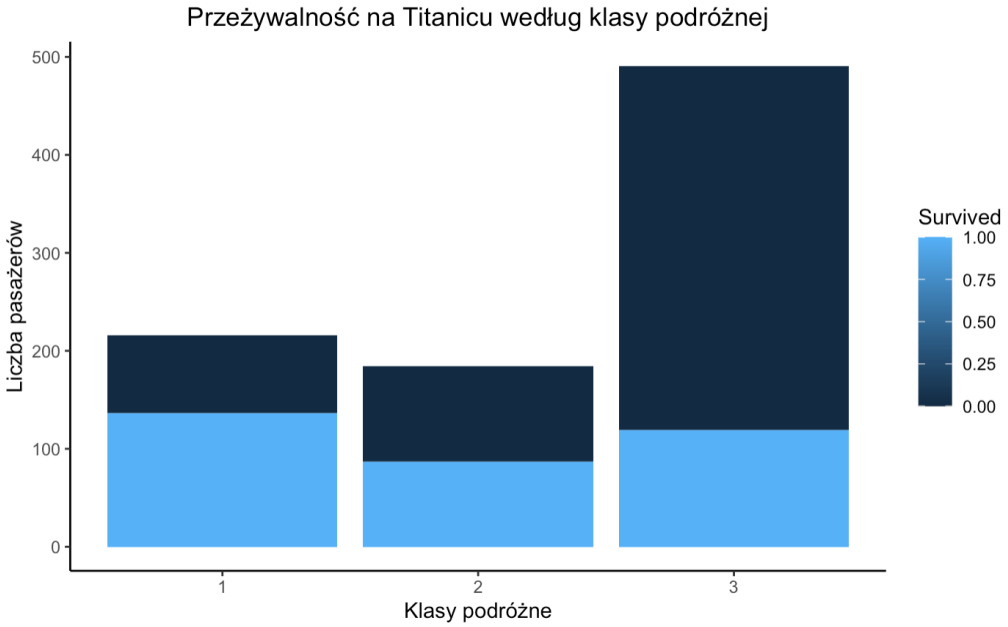
Na wykresie słupkowym można zaobserwować, że liczba mężczyzn na pokładzie Titanica była znacznie większa niż liczba kobiet, co potwierdza nasze wcześniejsze obliczenia i wnioski wynikające z opisu statystycznego danych. Wykres przedstawia jasną różnicę między liczebnością mężczyzn a kobiet na statku, gdzie liczba mężczyzn jest prawie dwukrotnie większa od liczby kobiet. Ta dysproporcja sugeruje, że Titanic był zdominowany przez mężczyzn pod względem liczby pasażerów.

Przyjrzyjmy się również liczbie ocalałych i zmarłych pasażerów. Na wykresie wyraźnie widać, że większość mężczyzn na pokładzie Titanica zginęła. Rzecz ma się przeciwnie, jeśli chodzi o kobiety.

Różnica w liczbie pasażerów na Titanicu, gdzie liczba mężczyzn była znacznie większa niż liczba kobiet, może wskazywać na istnienie pewnych społecznych i organizacyjnych wzorców tamtej epoki. W sytuacjach kryzysowych, takich jak katastrofa Titanica, priorytetem było zwykle ratowanie kobiet i dzieci. Dlatego też można wyjaśnić, dlaczego odsetek ocalałych kobiet był stosunkowo większy niż odsetek ocalałych mężczyzn. W tamtych czasach istniała pewna nieformalna zasada, że kobiety miały pierwszeństwo przy ewakuacji i dostępie do łodzi ratunkowych. Ta różnica w liczbie pasażerów i odsetku ocalałych między płciami może odzwierciedlać społeczne normy i wartości, które przeważały w tamtym okresie.

5.3 Przeżywalność na Titanicu według klasy podróżnej

Sprawdzimy, jak klasa, w której podróżowali pasażerowie, miała się do współczynnika ocalenia tych pasażerów. Czy pasażerowie w wyższych klasach mieli wyższy współczynnik ocalenia, niż ci w niższych klasach?

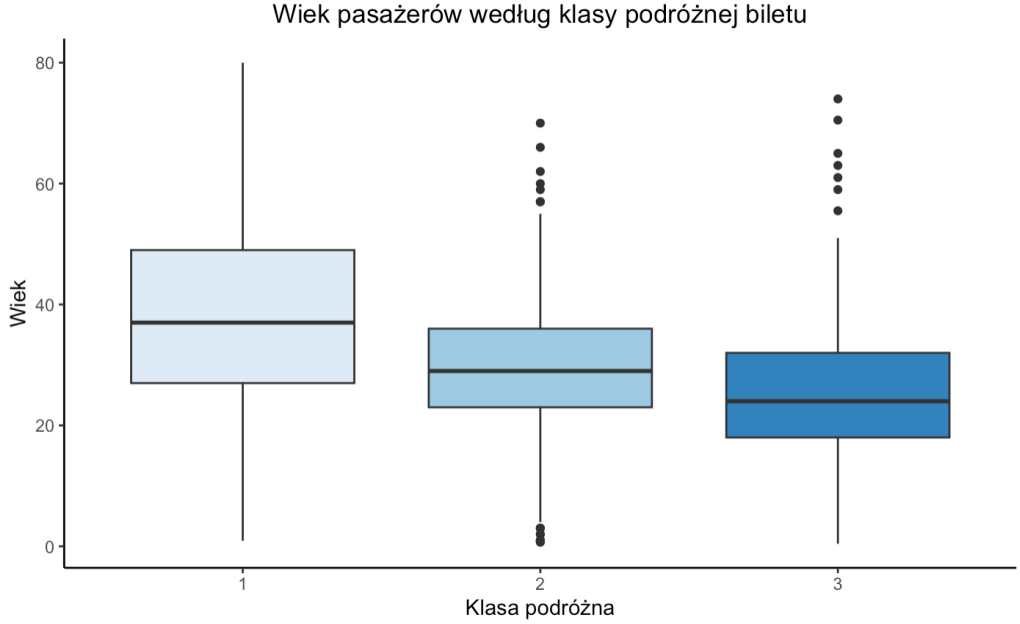


Powyższy wykres przedstawia zależność między klasą biletu (Pclass) a przeżyciem (Survived) pasażerów Titanica. Oś x reprezentuje klasę biletu, podczas gdy oś y przedstawia liczbę pasażerów. Kolor wypełnienia słupków wskazuje na przeżycie lub nieprzeżycie. Na wykresie możemy zaobserwować, jak przeżycie rozkładało się w różnych klasach biletów. Słupki w każdej klasie reprezentują liczbę pasażerów, a kolor wypełnienia wskazuje, czy dany pasażer przeżył (przez wypełnienie w jednym kolorze) czy nie przeżył (przez wypełnienie w innym kolorze).

Dzięki temu wykresowi możemy łatwo porównać liczbę pasażerów w poszczególnych klasach i zobaczyć, jak rozkładało się przeżycie w każdej z nich. Możemy również zauważyć, że klasa 1 miała największą liczbę ocalałych pasażerów, podczas gdy klasa 3 miała największą liczbę nieprzeżytych pasażerów. Powyższy wykres dostarcza wizualnego podsumowania danych dotyczących przeżycia w zależności od klasy biletu na Titanica.

5.4 Wiek pasażerów według klasy podróżnej

W kolejnym etapie naszej analizy skoncentrujemy się na badaniu związku między wiekiem a klasą podróżną. Naszym celem jest sprawdzenie, czy istnieje jakaś zależność pomiędzy wiekiem a klasą biletu. Czy pasażerowie podróżujący w wyższych klasach byli średnio starsi niż Ci w niższych klasach? Analiza tych informacji pomoże nam lepiej zrozumieć charakterystykę pasażerów podróżujących w różnych klasach.



Analizując wykres pudełkowy, możemy dostrzec różnice w wieku pasażerów w zależności od klasy podróżnej. Mediana wieku, czyli wartość środkowa, jest najwyższa w klasie 1, nieco niższa w klasie 2 i najniższa w klasie 3. Oznacza to, że przeciętny wiek pasażerów w poszczególnych klasach różni się od siebie. Pudełka na wykresie reprezentują zakres wieku, w którym znajduje się 50% danych. Im szersze pudełko, tym większy jest rozrzut wieku w danej klasie. W przypadku klasy 1 mamy największy zakres wieku, podczas gdy w klasach 2 i 3 jest nieco węższy.

Na wykresie widoczne są "wąsy", które wskazują rozpiętość danych, wyłączając wartości odstające. Możemy zauważyć, że klasa 1 ma największą różnicę wieku pomiędzy najmłodszą a najstarszą osobą, podczas gdy w klasach 2 i 3 ta różnica jest mniejsza. Punkty na wykresie reprezentują wartości odstające, czyli dane znajdujące się daleko od typowego rozkładu wieku w danej klasie. Możemy zobaczyć, że w klasie 2 i 3 istnieją pewne wartości odstające, które oznaczają pasażerów o znacznie wyższym wieku w porównaniu do reszty klasy.

Na podstawie tego wykresu pudełkowego możemy stwierdzić, że wiek pasażerów różni się w zależności od klasy podróżnej. Pasażerowie w klasie 1 są średnio starsi niż ci w klasach 2 i 3. Profil wiekowy poszczególnych klas podróżnych może mieć znaczenie w dalszej analizie i lepszym zrozumieniu wpływu wieku na inne zmienne w tym zbiorze danych.

6 Test parametryczny z interpretacją

Aby porównać różnicę w średnim wieku między grupą ocalałych i nieocalałych pasażerów, przeprowadzimy test t-Studenta, który jest testem parametrycznym. Nasze dane są niezależne, a wynik testu pozwoli nam ocenić, czy istnieje statystycznie istotna różnica między średnimi wieku w obu grupach.

```
        Welch Two Sample t-test

data:  Age by Survived
t = 2.046, df = 598.84, p-value = 0.04119
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 0.09158472 4.47339446
sample estimates:
mean in group 0 mean in group 1
       30.62618        28.34369
```

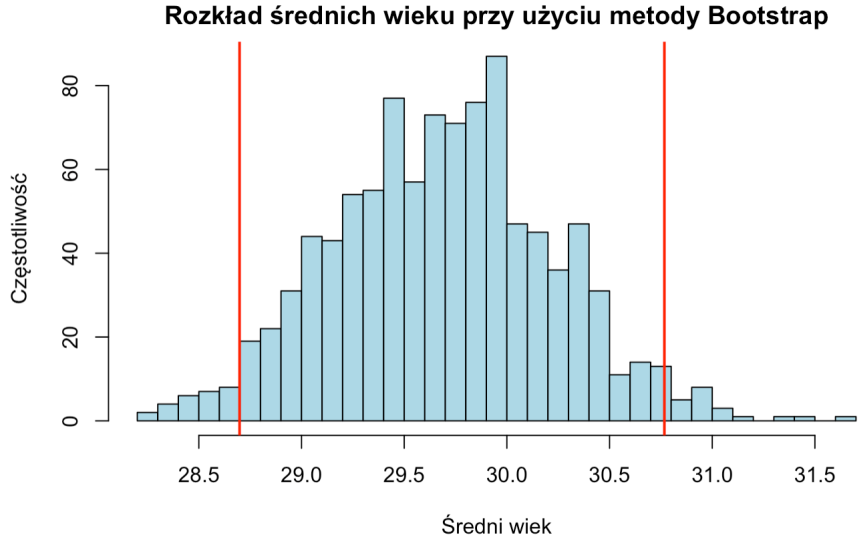
Wyniki testu t-Studenta wskazują, że wartość t-statystyki wynosi 2.046, a liczba stopni swobody wynosi około 598.84. P-value, która wynosi 0.04119, informuje nas, że istnieje statystycznie istotna różnica między średnimi wieku w grupie ocalałych i nieocalałych. To oznacza, że alternatywna hipoteza jest prawdziwa i istnieje rzeczywista różnica między średnimi wieku w tych dwóch grupach.

Przedział ufności 95%:[0.09158472, 4.47339446] wskazuje, że z 95% pewnością różnica między średnimi wieku w grupie nieocalałych i ocalałych mieści się w tym przedziale. Średni wiek w grupie nieocalałych wynosi 30.62618, podczas gdy w grupie ocalałych wynosi 28.34369.

Na podstawie tych wyników możemy stwierdzić, że istnieje statystycznie istotna różnica w wieku między osobami, które przeżyły i nie przeżyły katastrofy Titanica. Jednak różnica między średnimi wieku jest względnie mała, mieści się w przedziale ufności 95% i może nie mieć dużego znaczenia praktycznego.

7 Test nieparametryczny z interpretacją

Przeprowadziliśmy analizę metodą bootstrap w celu oszacowania przedziału ufności dla średniej wieku pasażerów Titanica. Metoda bootstrap polega na wielokrotnym losowaniu prób z powtórzeniami z naszych danych i obliczaniu średniej wieku dla każdej próby. Dzięki temu możemy uzyskać rozkład średnich wieku, który pomaga nam ocenić niepewność wokół estymatora średniej.



Na podstawie otrzymanego histogramu średnich wieku, możemy zauważyć, że rozkład tych średnich jest skoncentrowany wokół pewnej wartości, a różnice wokół tej wartości odzwierciedlają naturalną zmienność wyników przy losowaniu prób. Histogram przedstawia nam również rozkład tych średnich wieku, co pomaga nam wizualnie zobaczyć, jakie wartości są bardziej prawdopodobne. Otrzymany wynik przedziału ufności wskazuje, że średni wiek pasażerów Titanica,

oszacowany na podstawie metody bootstrap, mieści się w przedziale od 28.69819 do 30.76695. Oznacza to, że z prawdopodobieństwem 95% prawdziwa średnia wieku pasażerów Titanica znajduje się w tym zakresie. Widać, że oszacowana przez nas na początku pracy średnia wieku (29.7 lat) mieści się w wyznaczonym przy pomocy metody bootstrap przedziale.

Interpretując wynik, możemy stwierdzić, że na podstawie analizy danych próbkowych z metody bootstrap, średni wiek pasażerów Titanica wynosi około 29 lat. Przedział ufności pokazuje nam, że istnieje pewna niepewność wokół tej estymacji, ale z dużym prawdopodobieństwem możemy przyjąć, że rzeczywista średnia wieku pasażerów Titanica mieści się w tym zakresie.

Metoda bootstrap jest przydatnym narzędziem, gdy mamy ograniczoną liczbę próbek i chcemy ocenić niepewność naszych estymatorów. Przez wielokrotne losowanie prób i obliczanie statystyk interesujących nas, możemy uzyskać informacje o ich rozkładzie i przedziałach ufności. Jest to szczególnie przydatne w przypadku analizy danych, gdzie nie mamy dostępu do całej populacji i musimy polegać na próbkach, aby wnioskować o cechach populacji.

8 Podsumowanie i wnioski

Celem tego projektu było przeprowadzenie analizy danych pasażerów Titanica w celu lepszego zrozumienia różnych aspektów związanych z wiekiem, płcią, klasą podróżną oraz zależnościami między wiekiem a ceną biletu dla ocalałych pasażerów. Przy użyciu języka programowania R oraz biblioteki ggplot2, stworzyliśmy wizualizacje, które pomogły nam wizualnie przedstawić i wydedukować różne wzorce i trendy w danych. Na początku przeprowadziliśmy eksploracyjną

analizę danych, która obejmowała wczytanie i opisanie statystyczne danych. Zauważyliśmy, że większość pasażerów Titanica to mężczyźni, co sugeruje pewne wzorce społeczne i organizacyjne tamtej epoki, gdzie priorytetem było ratowanie kobiet i dzieci w sytuacjach kryzysowych.

Następnie skoncentrowaliśmy się na wieku pasażerów i stworzyliśmy histogram, który przedstawił rozkład wieku w różnych przedziałach wiekowych. Kolejnym krokiem było analizowanie wieku w zależności od klasy podróżnej. Wykorzystaliśmy wykres pudełkowy, który przedstawił rozkład wieku dla każdej z trzech klas podróżnych. Obserwowaliśmy różnice w medianach wieku między klasami. Przeprowadziliśmy także test t-Studenta oraz test nieparametryczny metodą bootstrap, aby porównać średni wiek między grupą ocalałych a nieocalałych pasażerów.

Podsumowując, analiza danych pasażerów Titanica dostarczyła nam cennych informacji na temat wieku, płci, klasy podróżnej i zależności między nimi. Zauważyliśmy różnice w wieku w zależności od płci i klasy podróżnej, co sugeruje, że demograficzny profil pasażerów był zróżnicowany. Pasażerowie w klasie 1 mieli tendencję do być starsi, podczas gdy w klasach 2 i 3 przeważała młodsza grupa wiekowa.

Analiza danych pasażerów Titanica pozwoliła nam również wnioskować o pewnych wzorcach społecznych i organizacyjnych tamtej epoki. Priorytetem było ratowanie kobiet i dzieci, co jest zgodne z obserwacjami dotyczącymi większej liczby ocalałych kobiet w porównaniu do mężczyzn. Również klasa podróżna wydaje się mieć wpływ na szanse przeżycia, gdzie pasażerowie z wyższych klas mieli większe szanse na ocalenie.

Warto podkreślić, że analiza danych pasażerów Titanica ma swoje ograniczenia. Dane, które mieliśmy do dyspozycji, nie były pełne, z pewnymi brakującymi wartościami. Ponadto, nasza analiza opierała się na dostępnych danym i zastosowanych metodach statystycznych. Dlatego wnioski, które wyciągnęliśmy, mają charakter obserwacyjny i nie można ich uważać za pełną reprezentację całej populacji pasażerów Titanica.