# *ANALIZA DANYCH DOTYCZĄCYCH PASAŻERÓW STATKU TITANIC*

*Kacper Andrzejewski 419925*

*W 1912r. wydarzyła się przejmująca tragedia w dziejach ludzkości. Największy ówczesny statek rejsowy świata zatonął, uderzając, niespodziewanie po 6 dniach podróży po Oceanie Atlantyckim, w górę lodową. Na pokładzie statku znajdowało się mnóstwo ludzi, w tym znanych osobistości i cenionych uczonych. Niestety jak się okazało po niedługiej chwili, większość podróżujących nie zdołała przeżyć. Na podstawie tego wydarzenia powstawało wiele teorii, analiz i innych treści takich jak np. film „Titanic". Z racji tego, iż znane jest wiele faktów i statystyk dotyczących pasażerów tego okrętu, można „pokusić się" o analizę i szukanie relacji między poszczególnymi zmiennymi.*

*W poniższym sprawozdaniu przedstawię analizę tego wydarzenia jak i szczegółów z nim związanych od strony statystyki. Przede wszystkim potrzebne do tego będą dane takie jak wiek, płeć, rodzaj klasy podróżniczej oraz informacja o przeżyciu pasażerów. Te zmienne pozwolą określić pewne relacje, przeprowadzić sprawniej analizę oraz wyciągnąć odpowiednie wnioski.*

*Do dyspozycji jest plik „titanic_new.csv" zawierający spis wszystkich ludzi znajdujących się wtedy na okręcie, oraz ich dane – kluczowe lub mniej istotne do przeprowadzenia analizy. Będę posługiwać się załączoną we wspomnianym pliku tabelą, jednakże zawiera ona pewne błędy lub niewprowadzone informacje w losowych miejscach, dlatego należy ją przygotować do obliczeń statystycznych, aby były one jak najbardziej rzetelne i poprawne. Z pomocą w tych działaniach przychodzi tzw. „Preprocessing", czyli wstępne przetwarzanie danych. Jest to to czyszczenie, przekształcanie oraz integrowanie służących do dalszych kroków. Dzięki temu dane ulegają poprawie, są wyższej jakości i umożliwiają stworzenie lepszych modeli w konsekwencji obliczeń.*

## *PREPROCESSING*

*Aby dokonać preprocessingu, czyli „wstępnej obróbki danych” w pierwszym kroku należy otworzyć plik „titanic_new.csv”. Znajdują się tam kolumny takie jak:*

*- Id pasażera*

*- informacja o przeżyciu katastrofy statku*

*- klasa podróżnicza pasażera*

*- imię i nazwisko danej osoby*

*- płeć pasażera*

*- wiek pasażera*

*- ilość rodzeństwa pasażera  na pokładzie*

*- ilość dzieci (potomstwa) pasażera na pokładzie*

*- nr biletu*

*- taryfa pasażerska*

*- nr kabiny (o ile pasażer ją posiadał)*

*- pierwsza litera portu brytyjskiego, z którego pasażer wypływał*

*Można z łatwością stwierdzić, że duża ilość tych danych jest nieistotna w analizie i ma charakter czysto poglądowy. Dlatego w pierwszym kroku preprocessingu usuwam kolumny, które zaznaczyłem na czerwono, gdyż nie zawierają one potrzebnych informacji w kontekście szukania relacji w danych pasażerów. Zostawiłem informacje o przeżyciu katastrofy ( 0 – nie, 1 – tak), klasa podróżnicza pasażera (1 – wysoka, 2- średnia, 3 – niska), płeć pasażera (Female – kobieta, Male – mężczyzna) oraz wiek pasażera (na statku znajdowały się osoby we wszystkich kategoriach wiekowych).*

*Przygotowana tabela musiała ulec kolejnym zmianom. Pewne komórki kolumny „wiek” były nieuzupełnione lub zawierały sprzeczne dane ( < 0 lub > 100). W związku z tym uzupełniłem brakujące dane poprzez policzenie mediany wieku wszystkich pasażerów, dzięki czemu nie*

*występują skrajne błędy statystyczne, gdyż mediana jest dobrym, oraz lepszym od średniej, wyznacznikiem, jeśli potrzebujemy znaleźć wartości średnie.*

*W kolejnym kroku należy przyjrzeć się kolumnie dotyczącej wskaźnika umieralności. W poszczególnych komórkach występowały pewnie anomalie, gdyż pojawiały się inne wartości niż „0„ – „nie przeżył" lub „1" – „przeżył". W przypadku duplikowania się cyfr 1 lub występowania innych np. „111" lub „432" można założyć, iż osoba przeżyła, a domyślnie miała tam występować wartość – w tym przypadku – 1. I na odwrót. Przy usuniętych, brakujących danych w kolumnie ciężko stwierdzić, czy dana osoba przeżyła czy nie, dlatego najbezpieczniej użyć innej metody preprocessingu – usuwania danych wierszy. Przy tak dużej liczbie osób usunięcie paru nie powinno wpłynąć znacząco na końcowe rezultaty.*

*W końcowym kroku należy zwrócić uwagę na kolumnę zawierającą komórki dotyczące płci. Są dwie opcje – female (kobieta) lub male (mężczyzna), natomiast czasami występowały tzw. „literówki" jak np. „Feemale" lub „ M4ale", które program rozpatruje oddzielnie, a człowiek jest w stanie odpowiednio zinterpretować. Dlatego poszczególne nieprawidłowe wyrazy należy zastąpić poprawnymi – female lub male. W ten sposób można doprowadzić do prawidłowego formatu kolumny „płeć – sex"*

*Przedstawione powyżej kroki doprowadziły mnie do klarownie i rzetelnie przygotowanej tabeli (FrameT1) do obliczeń statystycznych, analiz i tworzenia wniosków. Preprocessing jest istotny w kontekście takich badań, ponieważ poprawia on jakość danych dając w konsekwencji znacznie lepsze i dokładniejsze modele statystyczne.*

## *ANALIZA DANYCH STATYSTYCZNYCH*

*Na początku warto przedstawić podstawowe dane dotyczące ilości pasażerów w podziale na różne kategorie (zmienne):*

* *Ogólna ilość podróżujących: 891 – w tym:*
    - *ilość kobiet: 314,*
    - *ilość mężczyzn: 577.*
* *Ogólna ilość osób, która przeżyła: 342 – w tym:*
    - *ilość kobiet: 233,*
    - *ilość mężczyzn: 109.*

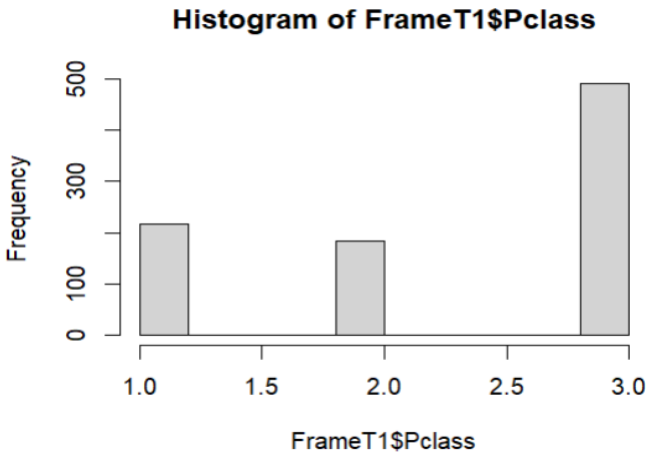
* *Ogólna ilość osób posiadających wykupione miejsce w kabinie: 204.*

* *Ogólna ilość osób z podziałem na klasy:*
    - *klasa 3: 491 pasażerów,*
    - *klasa 2: 184 pasażerów,*
    - *klasa 1: 216 pasażerów.*

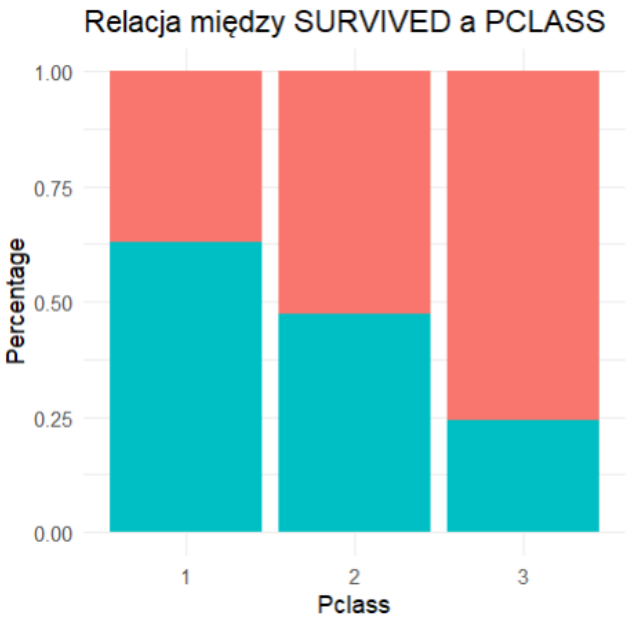
* *Ogólna ilość osób z podziałem na płeć:*
    - *0 – 20 lat: ok. 200,*
    - *20 – 40 lat: ok. 550,*
    - *40 – 60 lat: ok. 125,*
    - *60 – 80 lat: < 50.*

* *Średni wiek podróżujących: ~ 29.7 lat.*

**wszystkie dane zostały podane po preprocessingu**



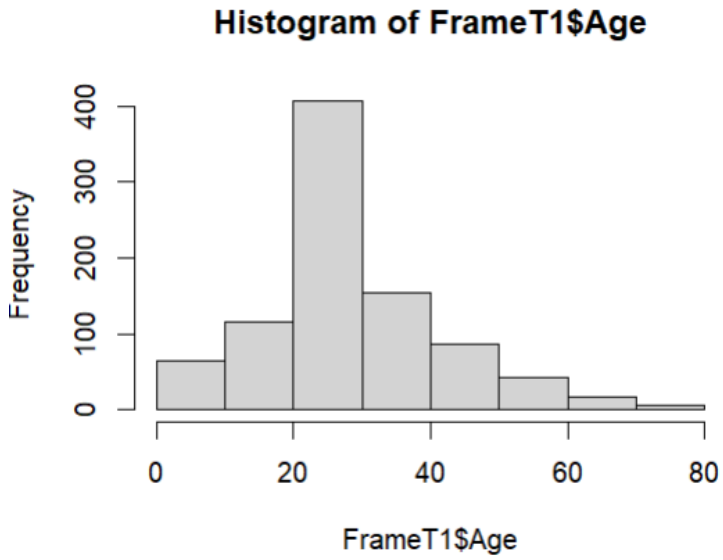

**** W pierwszym kroku przyjrzymy się wykresowi – histogramowi – obrazującemu zmienną „Pclass”, czyli rodzaj klasy podróżniczej. Najwięcej osób podróżowało klasą 3 – 491 pasażerów, co było spowodowane faktem, iż jest to najtańsza klasa, więc mogło na nią sobie statystycznie pozwolić najwięcej ludzi. Drugą klasą podróżowało najmniej osób, bo tylko 184, natomiast pierwszą – najwyższą – klasą niewiele więcej, bo 216, co jest zbliżone do wartości osób podróżujących 2 klasą. Mógł wpłynąć na to fakt, iż na okręcie znalazło się też wiele osobistości znanych, które mogły mieć w tamtych czasach większy status majątkowy, dlatego mogły sobie pozwolić na zakup biletu 1 i 2 klasy, 2 klasa mogła okazać się wystarczająca w udogodnieniach dla 184 osób, więc nie przekonywali się do wyższego komfortu.*




*Powyższy wykres przedstawia procentowy udział osób, które przeżyły katastrofę Titanica w poszczególnych klasach*

**** Z tego wykresu można wywnioskować, iż największą umieralność procentowo występowała w klasie 3, ponieważ znajdowała się tam największa ilość osób, natomiast warto zwrócić uwagę na pierwszą kolumnę. W 1 i 2 klasie znajdowała się podobna ilość osób, jak wspomniałem wyżej, natomiast w klasie 1 współczynnik umieralności jest niższy o ok. 20 punktów procentowych, co wskazuje na to, że podróżujący pierwszą klasą mieli prawdopodobnie lepsze zabezpieczenia przeciwwypadkowe, dzięki którym większy procent osób z nich przeżyło, porównując do pozostałych klas. Usprawiedliwia to przy okazji fakt z poprzedniej strony, iż największy odsetek zgonów był widoczny w 3 klasie, ponieważ miała ona właśnie najsłabsze zabezpieczenia.*

*** Przeanalizujmy teraz zmienną „age” oznaczającą wiek pasażerów. W tym celu posłużymy się histogramem prezentującym udział osób znajdujących się na statku w określonych kategoriach wiekowych:




*Jak widać na załączonym wykresie, zdecydowanie największy procent pasażerów stanowili ludzie w kategorii wiekowej 20 – 30 lat. Pozostali pasażerowie zajmowali coraz to mniejszy procent w ogólnej liczbie podróżujących. Lecz tak naprawdę ważniejszy jest wskaźnik prezentujący zależność między wiekiem a zmienną „Survived”, aby zauważyć jakąś prawidłowość między tymi zmiennymi. Dlatego warto obliczyć tym razem współczynnik korelacji między zmiennymi „Age” oraz „Survived”. Jest on równy -0.0370045, co jest wynikiem bardzo bliskim 0. Implikuje to fakt, iż występuje praktycznie brak zależności między wiekiem a przeżyciem. Oznacza to, że katastrofa statku niestety nie była tolerancyjna dla żadnej kategorii wiekowej, a sam wiek nie miał powiązania z tym, czy pasażer umrze, czy przeżyje.*

*** Spójrzmy teraz na korelację między wiekiem a wybieraną klasą podróżniczą. Może mieć to dosyć istotny wpływ na to, czy pasażer przeżył czy nie. Do tego celu posłużę się współczynnikiem korelacji metodą Spearmana (jest ona bardziej uniwersalna i tolerancyjna dla rozkładów nie-normalnych)*

```
> cor(FrameT1$Age, FrameT1$Pclass, method = 'spearman')
[1] -0.3174058
```

*Jak możemy zauważyć, współczynnik korelacji Spearmana między zmiennymi „Age" oraz „Pclass" nie jest zbyt bliski 0, co oznacza, że występuje pewna zależność między wiekiem a wybieraną klasą podróżniczą. Zapewne osoby starsze wybierały bardziej prestiżowe klasy, natomiast osoby młodsze mniej, ponieważ często podróżują z rodziną, dziećmi itd., więc nie mogą pozwolić sobie na zbyt duże wydatki. Jest to też powód tak dużej frekwencji w klasach niższych, a co za tym idzie wyższej umieralności w tej klasie (3).*

## *PODSUMOWANIE*

*Reasumując, po przeprowadzonej analizie powyższych zebranych danych, można dojść do konkluzji, iż na liczbę pasażerów, którzy (nie) przeżyli katastrofy/ę wpłynęło kilka istotnych czynników, takich jak np. wybrana klasa podróżnicza czy płeć. Do ogólnego wniosku mogą prowadzić różne ciągi przyczynowo-skutkowe złożone z takich danych jak np. większa ilość osób w tańszej klasie spowodowana podróżą z rodziną czy też bezpieczeństwo poszczególnych klas i ich zabezpieczenia wypadkowe. Podsumowując, była to bez wątpienia katastrofa na skalę światową, lecz dzięki dużej ilości zebranych danych oraz odpowiedniej analizie można wyciągnąć cenne wnioski oraz przyczyny takiego obrotu spraw. Preprocessing, analiza danych za*

*pomocą metod statystycznych oraz rozważne wyciąganie wniosków tworzą dobry i rzetelny model statystyczny, który może być cenny i ciekawy do zrozumienia natury tego wypadku. Osobiście uważam również, że przeprowadzona analiza może być czynnikiem kształtującym przyszłość, chociażby przyszłe i bezpieczne podróże statkiem. Oczywiście nikt nie umie przewidzieć tego, co się wydarzy, natomiast znacznie lepiej jest zapobiegać niż zwalczać.*