Wstęp do eksploracji danych

Raport z pracy domowej nr. 3

Katarzyna Mamla

Maj 2022

1 Treść pracy domowej

W ramach pracy domowej należało przeprowadzić eksperyment polegający na zbadaniu czy problem z czytaniem danych z różnych typów wykresów wciąż występuje w społeczeństwie.
Celem zadania było sprawdzenie czy „dobre praktyki" takie jak np. wybieranie słupków zamiast kątów czy unikanie wykresów 3D, faktycznie wpływają na poprawność informacji wydobywanych z wykresów.

2 Eksperyment

Istotą przeprowadzonego eksperymentu było sprawdzenie w jakim stopniu dane odczytwane z wykresu słupkowego 3D różnią od analogicznych danych odczytywanych z wykresu słupkowego 2D i jak często wystepują te rozbieżności.

2.1 Wykresy

Wykresy zostały przygotowane w oparciu o ramkę danych Titanic

PassengerId	Survived	Name	Sex	Age
1	0	Braund, Mr. Owen Harris	male	22.00
2	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00
3	1	Heikkinen, Miss. Laina	female	26.00
4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00
5	0	Allen, Mr. William Henry	male	35.00

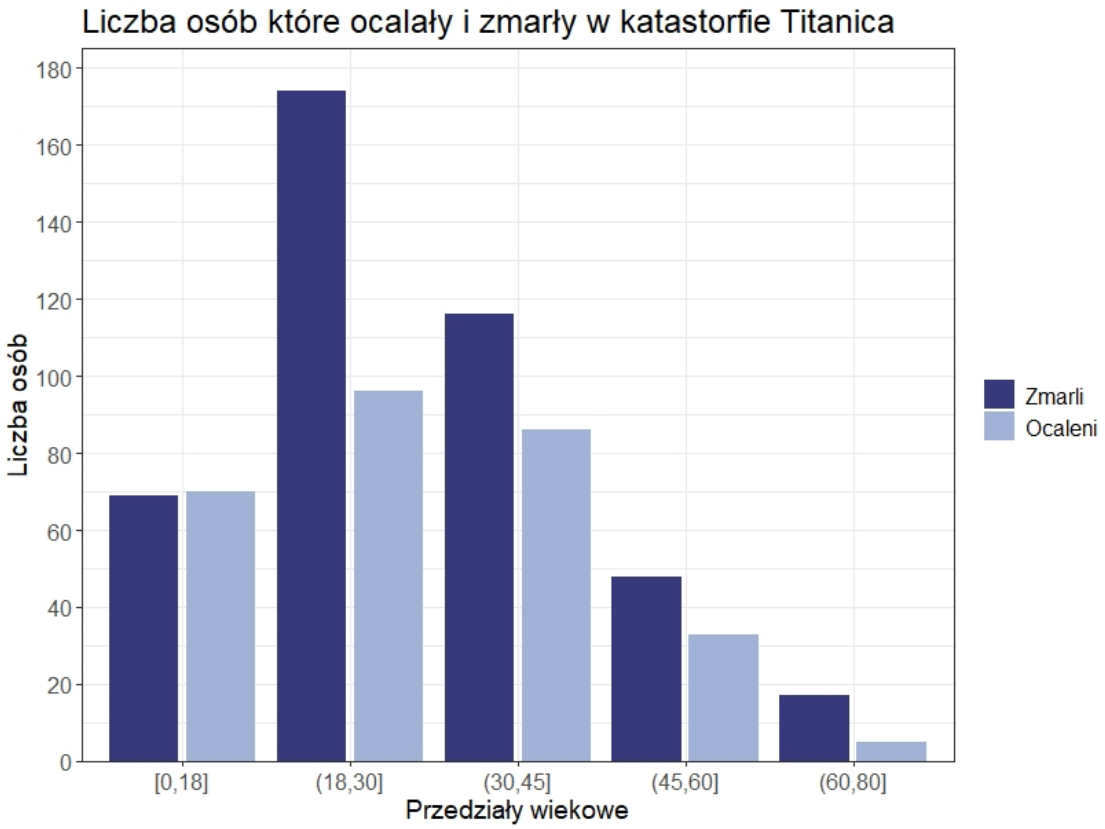
Tabela 1: Fragment ramki danych „Titanic"

Przedstawiają one jak, wśród poszczególnych grup wiekowych, rozkładała się liczba pasażerów, którzy ocaleli lub zmarli w katastrofie statku Titanic.

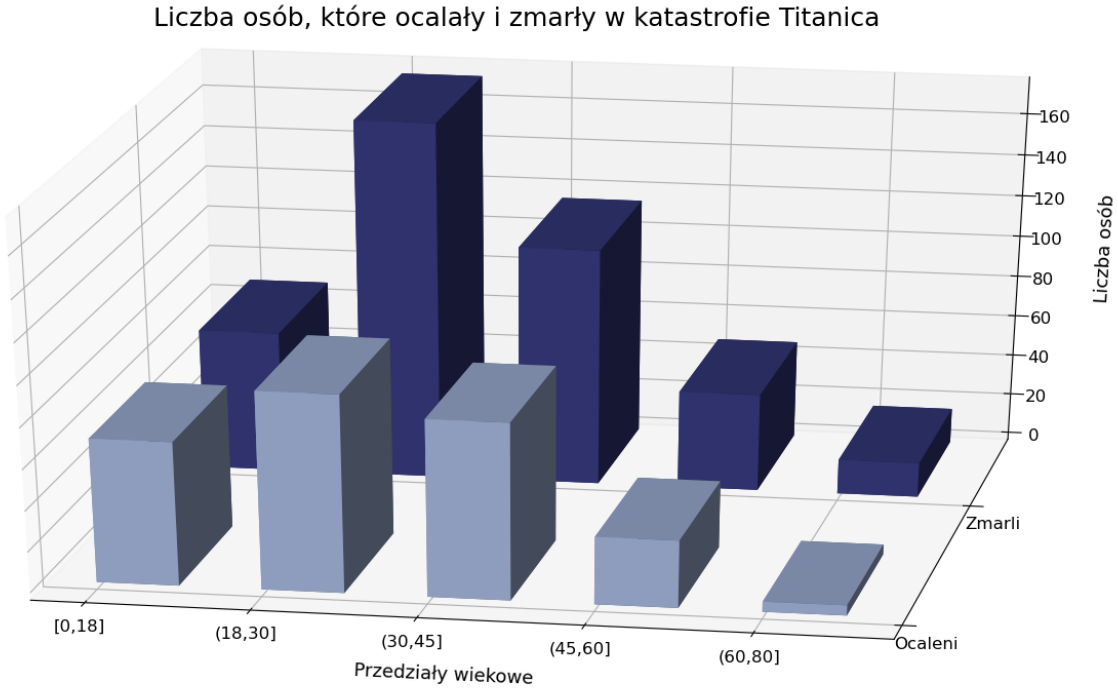
Kody do stworzenia wykresów znajdują się w Dodatku: kody 4

3 Ankieta

W celu zebrania informacji została przeprowadzona ankieta, w której wzieło udział 21 osób.
Ankietowani, na podstawie każdego z wykresów, proszeni byli o odczytanie lub oszacowanie następujących informacji:



Rysunek 1: Wykres 2D



Rysunek 2: Wykres 3D

- **P1**: Oceń, czy w przedziale wiekowym [0,18] więcej osób ocalało czy zmarło.
- **P2**: Oceń o ile więcej osób zmarło w przedziale wiekowym (30,45] niż ocalało w przedziale wiekowym (18,30].

A następnie o wybranie wykresu, który ich zdaniem był bardziej czytelny.

- **P3**: Który wykres był dla Ciebie bardziej czytelny?

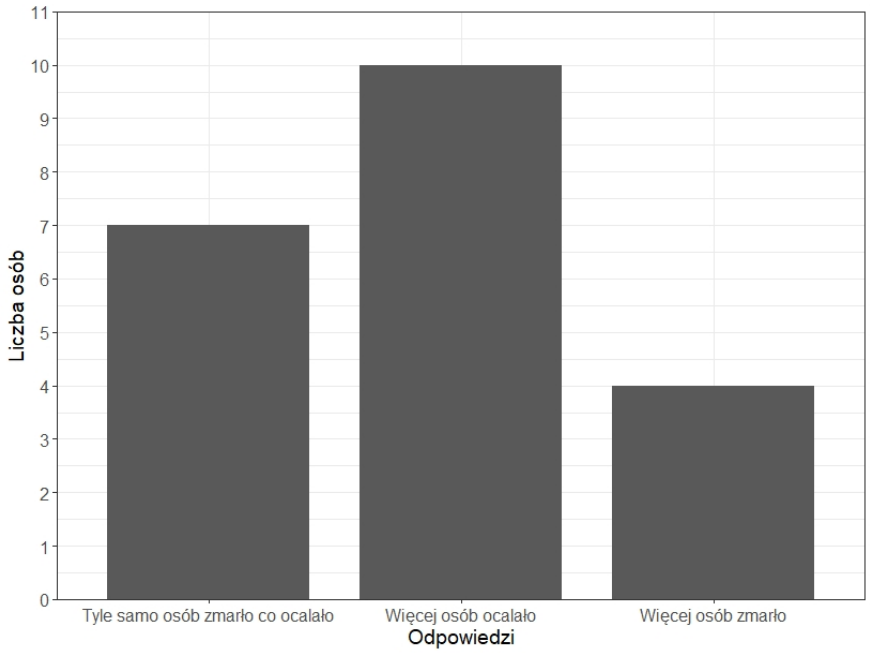
P1 3D	P1 2D	P2 3D	P2 2D	P3 Czytelność
Więcej ocalało	Więcej ocalało	5 - 10 os.	15 - 20 os.	Wykres 2D
Tyle samo zmarło co ocalało	Więcej ocalało	mniej niż 5 os.	15 - 20 os.	Wykres 2D
Tyle samo zmarło co ocalało	Więcej ocalało	5 - 10 os.	15 - 20 os.	Wykres 2D
Więcej ocalało	Więcej ocalało	5 - 10 os.	15 - 20 os.	Wykres 2D
Tyle samo zmarło co ocalało	Więcej ocalało	15 - 20 os.	15 - 20 os.	Wykres 2D
Tyle samo zmarło co ocalało	Więcej ocalało	5 - 10 os.	15 - 20 os.	Wykres 2D
Więcej ocalało	Więcej ocalało	15 - 20 os.	15 - 20 os.	Wykres 2D
Więcej ocalało	Więcej ocalało	więcej niż 20 os.	więcej niż 20 os.	Wykres 2D
Więcej ocalało	Więcej ocalało	15 - 20 os.	15 - 20 os.	Wykres 2D
Tyle samo zmarło co ocalało	Więcej ocalało	15 - 20 os.	15 - 20 os.	Wykres 2D
Więcej ocalało	Więcej ocalało	15 - 20 os.	15 - 20 os.	Wykres 2D
Tyle samo zmarło co ocalało	Więcej ocalało	5 - 10 os.	15 - 20 os.	Wykres 2D
Więcej zmarło	Więcej ocalało	5 - 10 os.	15 - 20 os.	Wykres 2D
Więcej ocalało	Więcej ocalało	5 - 10 os.	15 - 20 os.	Wykres 2D
Więcej zmarło	Więcej ocalało	więcej niż 20 os.	więcej niż 20 os.	Wykres 2D
Tyle samo osób zmarło co ocalało	Więcej ocalało	15 - 20 os.	15 - 20 osób	Wykres 2D
Więcej ocalało	Więcej ocalało	więcej niż 20 os.	więcej niż 20 os.	Wykres 2D
Więcej ocalało	Więcej ocalało	15 - 20 os.	15 - 20 os.	Wykres 2D
Więcej zmarło	Więcej ocalało	15 - 20 os.	15 - 20 os.	Wykres 2D
Więcej ocalało	Więcej ocalało	15 - 20 os.	15 - 20 os.	Wykres 2D
Więcej zmarło	Więcej ocalało	5 - 10 os.	15 - 20 os.	Wykres 2D

Tabela 2: Ogólne wyniki ankiety

3.1 Analiza wyników

3.1.1 Pytanie 1

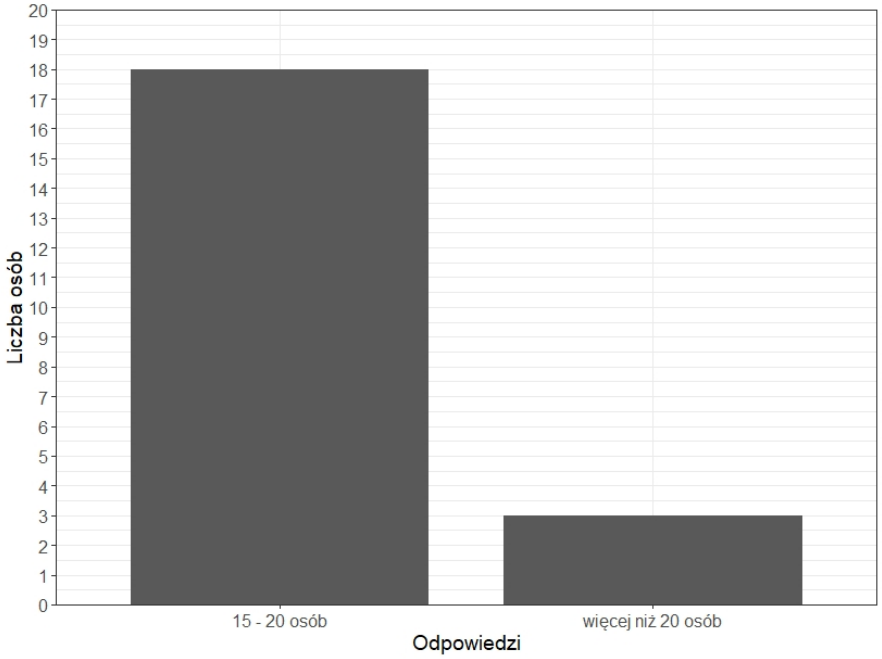
Na pytanie 1 (**P1**), w oparciu o wykres 2D, wszyscy respondenci odpowiedzieli poprawnie tzn. zaznaczyli odpowiedź *więcej osób ocalało*. Posługując się tym wykresem minimalna różnica danych tj. 1 osoba więcej wśród ocalałych jest łatwa do dostrzeżenia.
Odpowiedzi w oparciu o wykres 3D nie były jednoznaczne. Patrząc na Rysunek 3 można zauważyć, że najwięcej ankietowanych zaznaczyło poprawną odpowiedź, jest to jednak **niespełna 48% głosów**. Znaczna część osób zaznaczyła odpowiedź *Tyle samo osób zmarło co ocalało*. Minimalna różnica wśród ocalałych była tu więc cięższa do wychwycenia, a dla większości osób niewidoczna.



Rysunek 3: Rozkład głosów w pierwszym pytaniu dla wykresu 3D

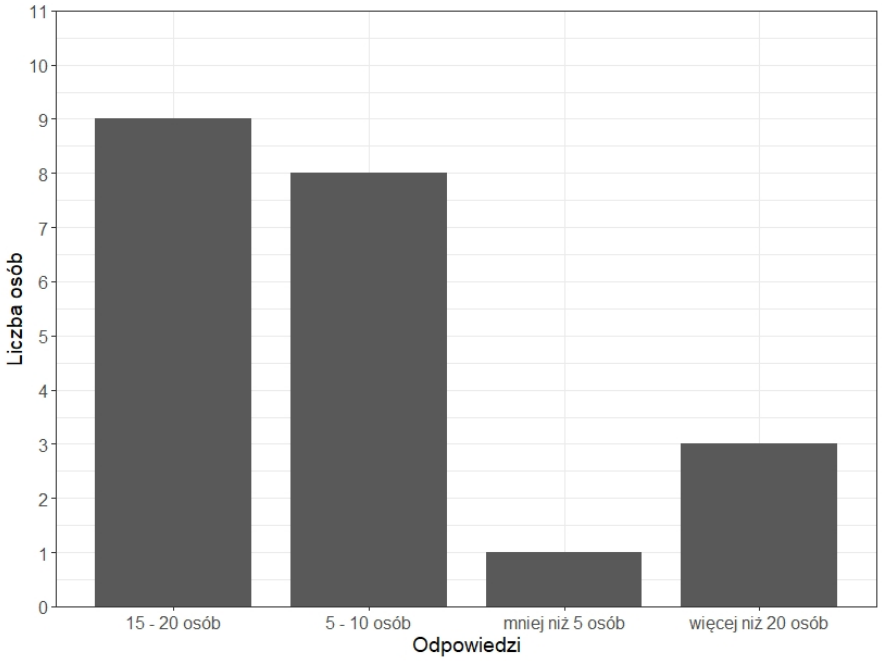
3.1.2 Pytanie 2

W oparciu o wykres 2D większość respondentów tj. 86% dobrze odpowiedziała na pytanie 2 (**P2**) tzn. zaznaczyła odpowiedź *15-20 osób*. Niewielka rozbieżność wśród głosów, widoczna na Rysunku 4, mogła wynikać z faktu, że poprawną odpowiedzią jest 20 osób.



Rysunek 4: Rozkdład głosów w pytaniu 2 dla wykresu 2D

Korzystając z wykresu 3D ankietowani odpowiadali najczęściej poprawnie, jednak ponownie jest to mniej niż połowa wszystkich osób tj. **43% głosów**. Drugą, co do ilości głosów, odpowiedzią była opcja *5-10 osób*, którą wybrało 38% osób. Można wnioskować, że porównywanie sporych różnic na wykresie 3D prowadzi do znacznych rozbieżności wśród danych odczytywanych przez różne osoby.



Rysunek 5: Rozkład głosów w pytaniu 2 dla wykresu 3D

3.1.3 Pytanie 3

Ankietowani jednogłośnie stwierdzili, że bardziej czytelny i przejrzysty jest wykres słupkowy 2D.

4 Wnioski

Dane odczytwane z wykresu słupkowego 3D znacząco różnią od analogicznych danych odczytywanych z wykresu słupkowego 2D. Ponad połowa badanych źle odczytała informacje z wykresu 3D, a nie miała problemów z analogiczną czynnością korzystając z wykresu 2D. Problem z czytaniem danych z wykresu 2D nie występuje, a dwuwymiarowe grafiki są preferowane i o wiele bardziej czytelne. Wykresy 3D nie są przeyjrzyste, a możliwość wyboru kąta patrzenia zaburza i znkiekształca prezentowane na nich wyniki.

Dodatek: kody

- Kod do stworzenia wykresu 2D (Rysunek 1) w *R*:

```
df %>%
  select(Survived,Age) %>%
  na.omit() %>%
  transmute(Survived, Intervals=cut(Age,c(0,18,30,45,60,80),
```

```
                                                labels=c("[0,18]",
                                                         "(18,30]",
                                                         "(30,45]",
                                                         "(45,60]",
                                                         "(60,80]"),
                                                         include.lowest= TRUE)) %>%
  group_by(Intervals) %>%
  count(Survived) -> tab1


  ggplot(tab1, aes(x = Intervals, y= n, fill=factor(Survived,
  labels= c("Zmarli", "Ocaleni")))) +
    geom_bar(stat="identity", width=.92, position = "dodge2")  +
    theme_bw()+
    scale_y_continuous(expand = c(0, 0),
                       limits = c(0, 185),
                       breaks = seq(0,180, by=20))+
    theme(legend.title = element_blank(),
          plot.title = element_text(size=18),
          axis.text=element_text(size=12),
          axis.title=element_text(size=14),
          legend.text = element_text(size=12))+
    labs(x= 'Przedzialy wiekowe',y='Liczba osob', fill=NULL)+
    scale_fill_manual(values = c("Zmarli" = "#36397B",
                                 "Ocaleni" = "#a1b2d7"))
```

- Kod do stworzenia wykresu 3D (Rysunek 2) w *Python*:

```
import matplotlib.pyplot as plt
import numpy as np
from matplotlib import rcParams
rcParams['axes.labelpad'] = 18
x = np.array(range(5), float)
y = np.array(range(2), float)
xpos, ypos = np.meshgrid(x,y)
z = np.array( [[70,96,86,33,5],[69,174,116,48,17]])
xpos=xpos.flatten()
ypos=ypos.flatten()
zpos= np.zeros_like(xpos)
dx= 0.5*np.ones_like(zpos)
dy= dx.copy()
dz =z.flatten()

fig= plt.figure()
ax= fig.add_subplot(111,projection ='3d')
ax.set_xticks(range(5))
xticks=['[0,18]','(18,30]','(30,45]','(45,60]','(60,80]']
ax.set_xticklabels(xticks, fontsize = 12)
ax.set_xlabel('Przedzialy wiekowe', fontsize = 13)
ax.set_yticks(range(2))
yticks=['Ocaleni','Zmarli']
ax.set_yticklabels(yticks, fontsize = 12)
colors=[]
colors.extend(["#a1b2d7"]*5)
```

```
colors.extend(['#36397B']*5)

ax.set_zticklabels([0, 20, 40,60,80, 100,120,140,160,180], fontsize= 12)
ax.set_zlabel('Liczba osob', fontsize = 13)


ax.bar3d(xpos,ypos,zpos,dx,dy,dz, color=colors)
plt.show()
```