Podstawy języka R - zadania (7 - dplyr i tidyr)

Tomasz Owczarek, Mateusz Naramski

2024/2025, semestr letni

Zbiór danych vgs.csv

Pod zmienną `vgs` wczytaj dane z pliku *vgs.csv*. Zamień kolumnę *genre* na zmienną czynnikową.

Zamiana na postać dłuższą (*pivot_longer*)

126. Przedstaw dane dotyczące 10 najlepiej sprzedających się gier PC na poszczególnych rynkach. Tabela z danymi powinna składać się z 3 kolumn: *name*, *market* (oznaczenie rynku) i *sales* (wielkość sprzedaży). W tym celu potokowo:

- przefiltruj dane wybierając tylko gry na platformę PC,
- wybierz kolumny *name*, *na_sales*, *eu_sales*, *jp_sales* i *other_sales*,
- zamień dane na postać dłuższą zbierając do jednej kolumny 2:5 (kolumna z nazwami ma nazywać się *market* a ta z wartościami - *sales*),
- posortuj (malejąco) wg kolumny *sales*,
- ogranicz wynik do 10 pierwszych rekordów.

```
## # A tibble: 10 x 3
##    name                                  market   sales
##    <chr>                                 <chr>    <dbl>
##  1 The Sims 3                            eu_sales  6.42
##  2 World of Warcraft                     eu_sales  6.21
##  3 Half-Life                             na_sales  4.03
##  4 Microsoft Flight Simulator            na_sales  3.22
##  5 Myst                                  eu_sales  2.79
##  6 World of Warcraft: The Burning Crusade na_sales  2.57
##  7 StarCraft II: Wings of Liberty        na_sales  2.56
##  8 Diablo III                            na_sales  2.43
##  9 Theme Hospital                        na_sales  2.3
## 10 Half-Life 2                           na_sales  2.28
```

127. Przedstaw w postaci długiej sprzedaż na rynkach amerykańskim, europejskim i japońskim gier platformowych (genre = Platform) z 1983 roku (nazwy kolumn mają być takie jak w rozwiązaniu poniżej).

```
## # A tibble: 15 x 3
##    name                     sales    value
##    <chr>                    <chr>    <dbl>
##  1 Mario Bros.              na_sales  0.51
##  2 Mario Bros.              eu_sales  0.12
##  3 Mario Bros.              jp_sales  1.63
##  4 Pitfall II: Lost Caverns na_sales  1.22
```

```
##  5 Pitfall II: Lost Caverns eu_sales  0.07
##  6 Pitfall II: Lost Caverns jp_sales  0
##  7 Donkey Kong              na_sales  0.23
##  8 Donkey Kong              eu_sales  0.05
##  9 Donkey Kong              jp_sales  0.84
## 10 Donkey Kong Jr.          na_sales  0.33
## 11 Donkey Kong Jr.          eu_sales  0.07
## 12 Donkey Kong Jr.          jp_sales  0.7
## 13 Popeye                   na_sales  0.51
## 14 Popeye                   eu_sales  0.12
## 15 Popeye                   jp_sales  0.45
```

Rozdzielanie kolumn (*separate*)

128. Zmodyfikuj rozwiązanie poprzedniego zadania tak żeby w kolumnie *market* znajdowały się wartości będące przed znakiem podkreślenia (tzn. *na*, *eu* lub *jp*). *Skorzystaj z funkcji separate i podziel kolumnę sales, jako nazwy rozdzielanych kolumn wskaż tylko jedną: "market".*

```
## # A tibble: 15 x 3
##    name                   market value
##    <chr>                  <chr>  <dbl>
##  1 Mario Bros.            na      0.51
##  2 Mario Bros.            eu      0.12
##  3 Mario Bros.            jp      1.63
##  4 Pitfall II: Lost Caverns na      1.22
##  5 Pitfall II: Lost Caverns eu      0.07
##  6 Pitfall II: Lost Caverns jp      0
##  7 Donkey Kong            na      0.23
##  8 Donkey Kong            eu      0.05
##  9 Donkey Kong            jp      0.84
## 10 Donkey Kong Jr.        na      0.33
## 11 Donkey Kong Jr.        eu      0.07
## 12 Donkey Kong Jr.        jp      0.7
## 13 Popeye                 na      0.51
## 14 Popeye                 eu      0.12
## 15 Popeye                 jp      0.45
```

Zamiana na postać szerszą (*pivot_wider*)

129. Przedstaw liczbę gier z poszczególnych gatunków wydanych w latach 2005-2010 w postaci szerokiej (wiersze mają wskazywać gatunki a kolumny lata). W tym celu potokowo:

- przefiltruj dane żeby wybrać tylko rekordy z odpowiednich lat,
- pogrupuj dane po roku i gatunku,
- podsumuj dane zliczając rekordy (funkcja `n()`),
- zamień wynik na postać szerszą wskazując rok jako kolumnę z nazwami (*names_from*), a kolumnę z liczbami rekordów jako kolumnę z wartościami (*values_from*).

```
## # A tibble: 12 x 7
##    genre       `2005` `2006` `2007` `2008` `2009` `2010`
##    <fct>        <int>  <int>  <int>  <int>  <int>  <int>
##  1 Action         192    184    211    221    272    226
##  2 Adventure       42     71     84    166    141    154
##  3 Fighting        43     55     50     57     53     40
##  4 Misc           115    109    151    212    207    201
```

```
##  5 Platform         83     54     42     62     29     31
##  6 Puzzle           33     43     66     64     79     45
##  7 Racing           77     75     86     82     84     57
##  8 Role-Playing     71    110    103    112    103    103
##  9 Shooter          96     69     85     83     91     81
## 10 Simulation       38     58     90    119    123     82
## 11 Sports          122    138    167    200    184    186
## 12 Strategy         29     42     67     50     65     53
```

130. Podobnie jak poprzednio przedstaw liczbę gier z poszczególnych gatunków wydanych w latach 2005-2010 w postaci szerokiej, ale tym razem to wiersze mają oznaczać lata, a kolumny gatunki (wyświetl tylko te gatunki, które w nazwie mają dużą lub małą literę "s"). *Skorzystaj z rozwiązanie poprzedniego zadania, tylko dodaj dodatkowe filtrowanie gatunku za pomocą funkcji* `grepl`*, a zamieniając na postać szeroką jako nazwy kolumn wskaż kolumnę genre*).

```
## # A tibble: 6 x 6
## # Groups:   year [6]
##    year  Misc Shooter Simulation Sports Strategy
##   <int> <int>   <int>      <int>  <int>    <int>
## 1  2005   115      96         38    122       29
## 2  2006   109      69         58    138       42
## 3  2007   151      85         90    167       67
## 4  2008   212      83        119    200       50
## 5  2009   207      91        123    184       65
## 6  2010   201      81         82    186       53
```

131. Wyświetl w postaci szerokiej średnią globalną sprzedaż (zaokrągloną do dwóch miejsc po przecinku) poszczególnych gatunków gier w latach 2012-2016 (wiersze to gatunki, kolumny to lata).

```
## # A tibble: 12 x 6
##    genre        `2012` `2013` `2014` `2015` `2016`
##    <fct>         <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Action         0.46   0.85   0.53   0.28   0.17
##  2 Adventure      0.1    0.11   0.08   0.15   0.05
##  3 Fighting       0.33   0.36   0.7    0.37   0.28
##  4 Misc           0.6    0.61   0.58   0.3    0.06
##  5 Platform       1.55   0.68   0.89   0.43   0.21
##  6 Puzzle         0.16   0.33   0.19   0.12  NA
##  7 Racing         0.48   0.82   0.62   0.42   0.08
##  8 Role-Playing   0.61   0.63   0.5    0.47   0.17
##  9 Shooter        1.52   1.06   1.4    1.95   0.57
## 10 Simulation     0.74   0.48   0.5    0.37   0.04
## 11 Sports         0.57   0.78   0.85   0.67   0.38
## 12 Strategy       0.22   0.33   0.12   0.11   0.05
```

Zbiór danych ufo.zip

Wczytaj i popraw dane z pliku *ufo.zip* wykonując poniższy kod:

```
# wczytanie danych
ufo <- read.csv(unzip("./data/ufo.zip", exdir = "./data"))

ufo$datetime <- mdy_hm(ufo$datetime)
ufo$date.posted <- mdy(ufo$date.posted)
ufo$country <- factor(ufo$country)
ufo$shape <- factor(ufo$shape)
```

132. Wyświetl liczbę przypadków zaobserwowanego UFO w latach 1941-1960 (*dołóż do ramki kolumnę z rokiem, pogrupuj po niej i policz rekordy*).

```
## # A tibble: 20 x 2
##     year     n
##    <dbl> <int>
##  1  1941     1
##  2  1942     2
##  3  1943     9
##  4  1944     9
##  5  1945     9
##  6  1946    10
##  7  1947    37
##  8  1948     8
##  9  1949    16
## 10  1950    28
## 11  1951    20
## 12  1952    49
## 13  1953    33
## 14  1954    53
## 15  1955    32
## 16  1956    45
## 17  1957    72
## 18  1958    47
## 19  1959    50
## 20  1960    66
```

133. Wyświetl liczbę przypadków zaobserwowanego UFO o kształtach “light” i “disk” w latach 1994-2002.

```
## # A tibble: 18 x 3
## # Groups:   year [9]
##     year shape     n
##    <dbl> <fct> <int>
##  1  1994 disk     47
##  2  1994 light    61
##  3  1995 disk     42
##  4  1995 light    97
##  5  1996 disk     55
##  6  1996 light   120
##  7  1997 disk     84
##  8  1997 light   274
##  9  1998 disk    107
```

```
## 10  1998 light   327
## 11  1999 disk    168
## 12  1999 light   539
## 13  2000 disk    179
## 14  2000 light   557
## 15  2001 disk    187
## 16  2001 light   673
## 17  2002 disk    216
## 18  2002 light   676
```

134. Przedstaw to samo, co w zadaniu **133.**, ale w postaci szerokiej (liczba "świateł" i "dysków" w osobnych kolumnach).

```
## # A tibble: 9 x 3
## # Groups:   year [9]
##    year  disk light
##   <dbl> <int> <int>
## 1  1994    47    61
## 2  1995    42    97
## 3  1996    55   120
## 4  1997    84   274
## 5  1998   107   327
## 6  1999   168   539
## 7  2000   179   557
## 8  2001   187   673
## 9  2002   216   676
```

135. Utwórz tabelę w szerokiej postaci, zawierającą liczbę obserwacji UFO o kształcie dysku, kuli ognia oraz trójkąta, w stanach Floryda (fl), Nowy York (ny) i Texas (tx).

```
## # A tibble: 3 x 4
##   shape       fl    ny    tx
##   <fct>    <int> <int> <int>
## 1 disk       242   239   238
## 2 fireball   452   240   215
## 3 triangle   385   314   382
```

136. Wyświetl pierwszych 20 rekordów ze zbioru `ufo`, rodzielając kolumnę `datetime` na dwie: date oraz time, zawierające odpowiednio datę i godzinę obserwacji. Poza datą wyświetl też kształt zaobserwowanego UFO.

```
##          date     time    shape
## 1  1949-10-10 20:30:00 cylinder
## 2  1949-10-10 21:00:00    light
## 3  1955-10-10 17:00:00   circle
## 4  1956-10-10 21:00:00   circle
## 5  1960-10-10 20:00:00    light
## 6  1961-10-10 19:00:00   sphere
## 7  1965-10-10 21:00:00   circle
## 8  1965-10-10 23:45:00     disk
## 9  1966-10-10 20:00:00     disk
## 10 1966-10-10 21:00:00     disk
## 11 1968-10-10 13:00:00   circle
## 12 1968-10-10 19:00:00 fireball
## 13 1970-10-10 16:00:00     disk
```

```
## 14 1970-10-10 19:00:00  unknown
## 15 1971-10-10 21:00:00     oval
## 16 1972-10-10 19:00:00   circle
## 17 1972-10-10 22:30:00     disk
## 18 1973-10-10 19:00:00     disk
## 19 1973-10-10 23:00:00    light
## 20 1974-10-10 19:30:00    other
```

137. Sprawdź, w których godzinach doby (bez minut) najczęściej było widziane UFO. Wyniki uporządkuj wg liczby wystąpień UFO.

Pierwszych 10 rekordów:

```
## # A tibble: 10 x 2
##       hr     n
##    <int> <int>
##  1    21 11445
##  2    22 10837
##  3    20  8617
##  4    23  7953
##  5    19  6147
##  6     0  4802
##  7    18  4002
##  8     1  3210
##  9    17  2592
## 10     2  2357
```

138. Wyświetl liczbę obserwacji poszczególnych kształtów UFO, ale tylko w XXI wieku. Ogranicz wynik tylko do tych kształtów, które widziane były przynajmniej 1000 razy, a wyniki uporządkuj malejąco wg liczby obserwacji.

```
## # A tibble: 13 x 2
##    shape         n
##    <fct>     <int>
##  1 light     13720
##  2 circle     6189
##  3 triangle   5916
##  4 fireball   5124
##  5 unknown    4511
##  6 sphere     4235
##  7 other      4204
##  8 disk       3097
##  9 oval       2865
## 10 formation  1995
## 11 changing   1623
## 12 cigar      1404
## 13 flash      1137
```

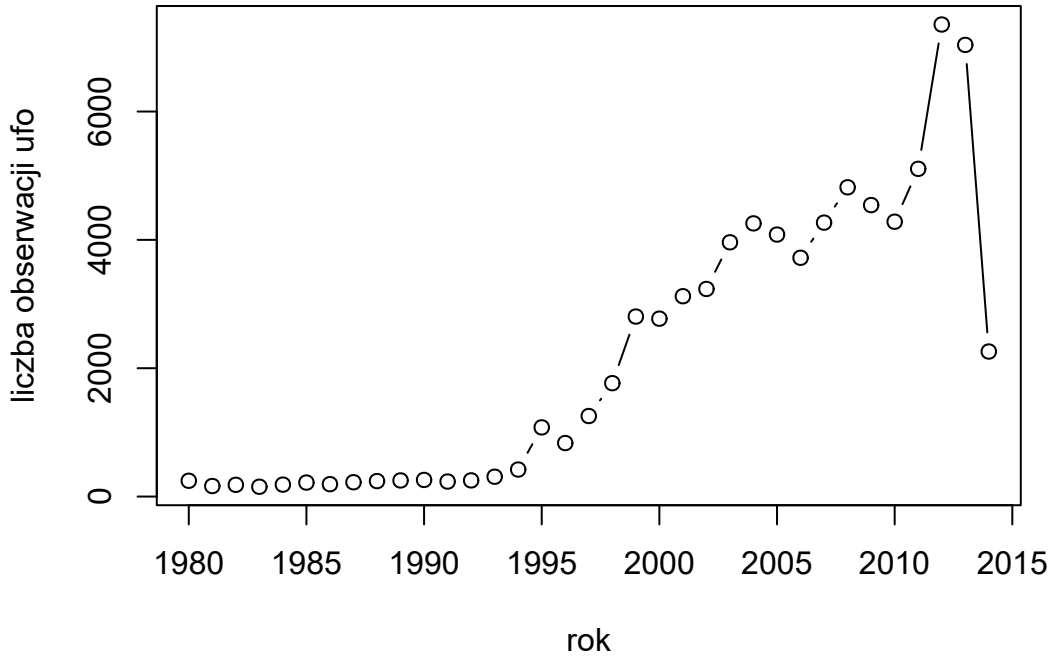
139. Wyświetl liczbę obserwacji poszczególnych kształtów ufo z podziałem na ich zaobserwowanie w XX i w XXI wieku. Uporządkuj wyniki wg sumarycznej łącznej obserwacji każdego kształtu (malejąco).

Pierwszych 10 rekordów:

```
## # A tibble: 10 x 3
##    shape     `XX w.` `XXI w.`
##    <fct>       <int>    <int>
```

```
##  1 light        2845    13720
##  2 triangle     1949     5916
##  3 circle       1419     6189
##  4 fireball     1084     5124
##  5 other        1445     4204
##  6 unknown      1073     4511
##  7 sphere       1152     4235
##  8 disk         2116     3097
##  9 oval          868     2865
## 10 formation     462     1995
```

140. Utwórz poniższy wykres.



Piraci znowu w ataku!

Wykonaj poniższy kod ładujący pakiet `yarrr` i podstawiający pod zmienna `pir` dane o piratach.

```
library(yarrr)
pir <- pirates
pir$sex <- factor(pir$sex)
pir$headband <- factor(pir$headband)
pir$college <- factor(pir$college)
pir$favorite.pirate <- factor(pir$favorite.pirate)
pir$sword.type <- factor(pir$sword.type)
pir$fav.pixar <- factor(pir$fav.pixar)
```

141. Wyświetl średnią liczbę zdobytych skrzynek ze skarbami w podziale na płeć piratów i uzywaną przez nich broń.

```
## # A tibble: 12 x 3
## # Groups:   sword.type [4]
##    sword.type sex    mean.tchests
##    <fct>      <fct>         <dbl>
##  1 banana     female         30.4
##  2 banana     male           15.5
##  3 banana     other          11.3
##  4 cutlass    female         25.6
##  5 cutlass    male           20.9
##  6 cutlass    other          29.0
##  7 sabre      female         19.4
##  8 sabre      male           15.7
##  9 sabre      other          20.8
## 10 scimitar   female         20.2
## 11 scimitar   male           16.9
## 12 scimitar   other          15
```

142. Dane z poprzedniego zadania przedstaw w postaci szerokiej.

```
## # A tibble: 4 x 4
## # Groups:   sword.type [4]
##   sword.type female  male other
##   <fct>       <dbl> <dbl> <dbl>
## 1 banana       30.4  15.5  11.3
## 2 cutlass      25.6  20.9  29.0
## 3 sabre        19.4  15.7  20.8
## 4 scimitar     20.2  16.9  15
```

143. Wyświetl w postaci szerokiej średni wiek piratów poszczególnych płci w podziale na uczelnie, które ukończyli.

```
## # A tibble: 2 x 4
## # Groups:   college [2]
##   college female  male other
##   <fct>    <dbl> <dbl> <dbl>
## 1 CCCC      25.9  23.3  24.8
## 2 JSSFP     33.8  32.0  32.5
```

144. Wyświetl liczbę, średni wiek i średnią liczbę skrzynek zdobytych przez absolwentów poszczególnych uczelni każdej płci.

```
## # A tibble: 6 x 5
## # Groups:   college [2]
##   college sex        n mean.age mean.tchests
##   <fct>   <fct>  <int>    <dbl>        <dbl>
## 1 CCCC    female   228     25.9         21.4
## 2 CCCC    male     397     23.3         18.2
## 3 CCCC    other     33     24.8         21.0
## 4 JSSFP   female   236     33.8         28.5
## 5 JSSFP   male      93     32.0         28.4
## 6 JSSFP   other     13     32.5         40.5
```

145. Dane z poprzedniego zadania przedstaw postaci długiej.

Pierwszych 12 wierszy:

```
## # A tibble: 12 x 4
## # Groups:   college [2]
##    college sex    name         value
##    <fct>   <fct>  <chr>        <dbl>
##  1 CCCC    female n            228
##  2 CCCC    female mean.age      25.9
##  3 CCCC    female mean.tchests  21.4
##  4 CCCC    male   n            397
##  5 CCCC    male   mean.age      23.3
##  6 CCCC    male   mean.tchests  18.2
##  7 CCCC    other  n             33
##  8 CCCC    other  mean.age      24.8
##  9 CCCC    other  mean.tchests  21.0
## 10 JSSFP   female n            236
## 11 JSSFP   female mean.age      33.8
## 12 JSSFP   female mean.tchests  28.5
```

146. Dla każdego znanego pirata wyświetl liczbę piratów każdej płci, wśród których są ulubionymi piratami.

```
## # A tibble: 6 x 4
## # Groups:   favorite.pirate [6]
##   favorite.pirate female  male other
##   <fct>            <int> <int> <int>
## 1 Anicetus            59    20    10
## 2 Blackbeard          87    23     9
## 3 Edward Low          71    24     7
## 4 Hook                85    23     9
## 5 Jack Sparrow        71   379     5
## 6 Lewis Scot          91    21     6
```

147. To samo, co w poprzednim zadaniu, ale wyniki przedstaw procentowo dla każdej płci (*wskazówka pod koniec materiałów z* `dplyr`).

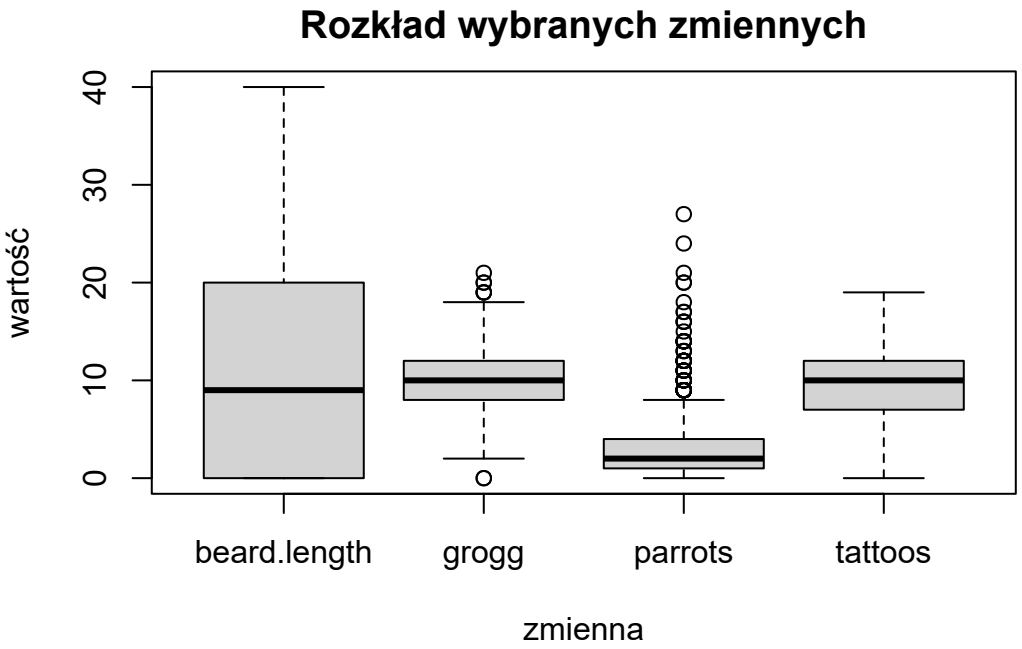
```
## # A tibble: 6 x 4
## # Groups:   favorite.pirate [6]
##   favorite.pirate female  male other
##   <fct>            <dbl> <dbl> <dbl>
## 1 Anicetus          12.7  4.08  21.7
## 2 Blackbeard        18.8  4.69  19.6
```

```
## 3 Edward Low        15.3  4.90  15.2
## 4 Hook              18.3  4.69  19.6
## 5 Jack Sparrow      15.3 77.3   10.9
## 6 Lewis Scot        19.6  4.29  13.0
```

148. Wyświetl ulubione filmy Pixara i liczbę piratów każdej płci, które je lubią (sprawdź w pomocy funkcji `pivot_wider`, który argument pozwala wypełniać brakujące wartości, żeby zamiast `NA` było 0).

```
## # A tibble: 15 x 4
## # Groups:   fav.pixar [15]
##    fav.pixar           female  male other
##    <fct>                <int> <int> <int>
##  1 A Bug's Life            21     6     1
##  2 Brave                    7    11     3
##  3 Cars                    10     9     2
##  4 Cars 2                   9     5     0
##  5 Finding Nemo            67    85    10
##  6 Inside Out             139   154    11
##  7 Monsters University     43    48     3
##  8 Monsters, Inc.          20    28     2
##  9 Ratatouille              1     1     1
## 10 The Incredibles         34    28     4
## 11 Toy Story                6     9     1
## 12 Toy Story 2             10    14     1
## 13 Toy Story 3              7     6     0
## 14 Up                      55    59     4
## 15 WALL-E                  35    27     3
```

149. Wyświetl poniższy wykres (wskazówka: utwórz ramkę z odpowiednimi zmiennymi w postaci długiej i skorzystaj z niej przy tworzeniu wykresu).



Zadanie dodatkowe

150! W pliku *wyniki_studentow.csv* znajdują się wyniki studentów A:Z z dziewięciu testów. Wczytaj ten plik a następnie utwórz listę rankingową studentów. Lista powinna zawierać oznaczenie studenta, jego wyniki z poszczególnych testów oraz jego średnią ze wszystkich testów. Lista powinna być uporządkowana wg średniej malejąco. Cztery pierwsze rekordy poniżej.

```
##   student test1 test2 test3 test4 test5 test6 test7 test8 test9 srednia
## 1       C     4     9     9     7     8     4     5     8     9    7.00
## 2       W     9     8     3     6     7     9     2     7     8    6.56
## 3       X    10     3     9     8     7     8     5     1     8    6.56
## 4       M     7     5     8     9     2     6     8    10     3    6.44
```