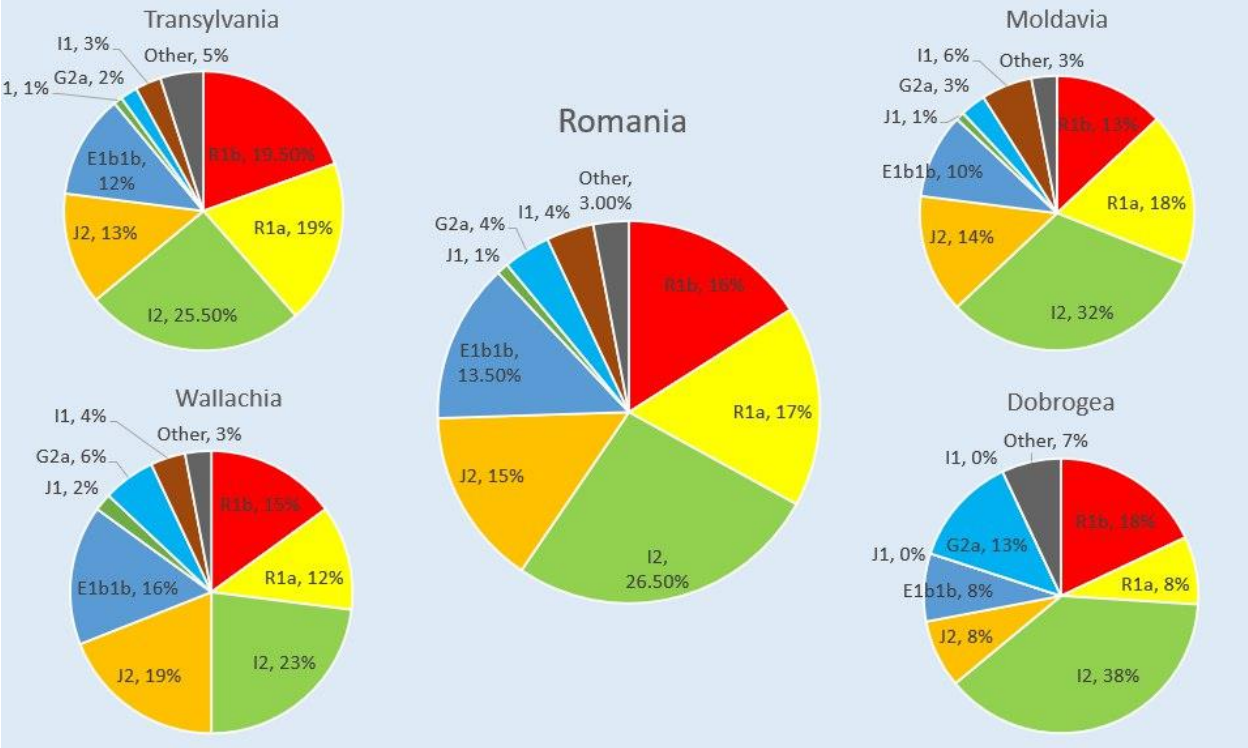
Praca domowa 3

W ramach pracy domowej № 3 rozważyłam wizualizację danych ze strony https://en.m.wikipedia.org/wiki/Romanians pokazującą odsetek populacji Rumunii i jej regionów mający tą lub inną haplogrupę DNA chromosomu Y:



Wykres ten został opublikowany 27 marca 2024 roku, co można sprawdzić na stronie https://commons.wikimedia.org/wiki/File:Romanian_Y-dna_Pie_Charts.png#mw-jump-to-license.

Łatwo zauważyć, że powyższa wizualizacja nie jest najbardziej skutecznym przedstawieniem danych, ponieważ ludzie mają słabą zdolność porównywania pól wycinków koła. Oprócz tego, skoro wykres środkowy jest większy od pozostałych, trudno jest porównac dane na tym wykresie ze wszystkimi innymi.

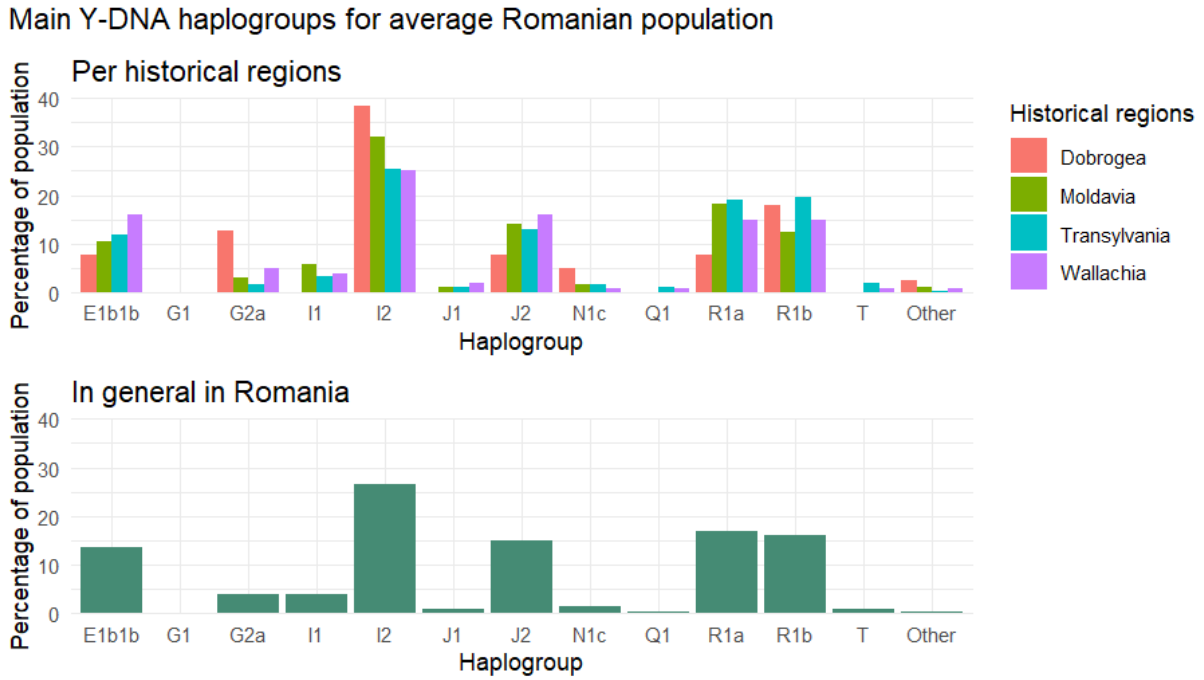
Najłatwieszym sposobem na poprawę tej wizualizacji jest stworzenie zamiast wykresu kołowego wykresu słupkowego. W tym celu skorzystałam z danych ze strony https://docs.google.com/spreadsheets/d/1Oc6XHFXRaZI4LBs28q4NnESRdCs8_35M8NNxZ5kyhKU/edit?pli=1#gid=783122462 i stworzyłam na ich podstawie ramkę danych `haplogroups.xlsx` (można ją zobaczyć na stronie https://docs.google.com/spreadsheets/d/1kzz0K-LVFVqlweCznLczwU4A607DVt23OK2TV1lwyEU/edit?usp=sharing):

Haplogroup	Dobrogea	Transylvania	Moldavia	Wallachia	Romania
R1b	18%	19,50%	13%	15%	16,00%
R1a	8%	19%	18%	15%	17,00%
I2	38%	25,50%	32%	25%	26,50%
J2	8%	13%	14%	16%	15,00%
E1b1b	8%	12%	10%	16%	13,50%
J1	0%	1%	1%	2%	1,00%
G2a	13%	2%	3%	5%	4,00%
I1	0%	3%	6%	4%	4,00%
N1c	5%	2%	2%	1%	1,50%
T	0%	2%	0%	1%	1,00%
G1	0%	0%	0%	0%	0,00%
Q1	0%	1%	0%	1%	0,50%
E (other)	2%	0%	0%	1%	0,50%
Other	0%	0%	1%	0%	0,00%

Implementacja:

```
install.packages("readxl")

install.packages("patchwork")

library(readxl)

library(dplyr)

library(tidyverse)

library(ggplot2)

library(patchwork)


haplogroups <- read_excel("haplogroups.xlsx")

haplogroups[14, -1] <- haplogroups[13, -1] + haplogroups[14, -1]

haplogroups <- haplogroups[-13, ] %>% pivot_longer(!Haplogroup, names_to = "Region",
values_to = "Percentage")

haplogroups$Haplogroup <- factor(haplogroups$Haplogroup,

                                 levels = c("E1b1b", "G1", "G2a", "I1", "I2",

                                            "J1", "J2", "N1c", "Q1", "R1a",

                                            "R1b", "T", "Other"))

regions_plot <- haplogroups %>%
```

```
  filter(Region != "Romania") %>%
  ggplot(aes(x = Haplogroup, y = Percentage*100, fill = Region)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_y_continuous(expand = c(0, 0), limits = c(0, 40)) +
  labs(title = "Per historical regions", x = "Haplogroup",
       y = "Percentage of population", fill = "Historical regions") +
  theme_minimal()
romania_plot <- haplogroups %>%
  filter(Region == "Romania") %>%
  ggplot(aes(x = Haplogroup, y = Percentage*100)) +
  geom_bar(stat = "identity", fill = "aquamarine4") +
  scale_y_continuous(expand = c(0, 0), limits = c(0, 40)) +
  labs(title = "In general in Romania",
       x = "Haplogroup", y = "Percentage of population") +
  theme_minimal()
regions_plot / romania_plot +
  plot_annotation(title = "Main Y-DNA haplogroups for average Romanian population")
```



Powyższa wizualizacja pozwala łatwo zauważyć najczęstszą i najrzadszą haplogrupę i porównać te wyniki w zależności od regionu. Ponadto wykres danych z całego kraju ma taką samą skalę i takie same oznaczenia, jak i wykres danych z poszczególnych regionów, dzięki czemu możemy skutecznie porównać te dwa wykresy.