WYDZIAŁ
MATEMATYKI
I FIZYKI STOSOWANEJ
POLITECHNIKI RZESZOWSKIEJ

DZIENNIK

PROJEKT GRUPOWY

Miliarderzy według Forbes'a - Analiza Danych

Wykonali:
Gabriela Kiwacka
Piotr Greń

Inżynieria i Analiza Danych
Grupa laboratoryjna nr 2

Opiekun pracy:
dr. Ewa Rejwer – Kosińska

Rzeszów, 2022

Spis treści

.....2
Opis projektu.....3
Etapy tworzenia projektu.....3
Dzień 1.....3
Dzień 2.....5
Dzień 3.....6
Dzień 4.....8
Dzień 5.....10
Dzień 6.....11
Dzień 7.....12
Dzień 8.....13
Dzień 9.....15
Wnioski.....15
Źródła:.....15

Opis projektu

Projekt będzie skupiał się na analizie danych na temat miliarderów zgromadzonych przez magazyn Forbes. Posługiwać będziemy się ramką danych „billionaires” gromadzącą informacje na temat 2600 miliarderów na przestrzeni lat: 2015 – 2022. Ramka danych nie była w pełni gotowa, a mianowicie dane z lat 2015 – 2021 były do niej dołączane z innych ramek. Miliarderzy w ramce danych posortowani są według rankingu od najbogatszego. Ponadto ramka danych zawiera takie dane jak: imię i nazwisko, wiek miliardera, kraj pochodzenia, źródło dochodu, firma, w której pracują oraz w jakiej branży. Jednak najbardziej istotnymi danymi jakie możemy znaleźć w tej ramce danych są informacje na temat wysokości majątku w każdym roku z wyżej wymienionego zakresu. W projekcie będziemy zajmować się analizą poszczególnych danych, które będą przedstawiane w plikach, a na ich podstawie będą tworzone wykresy np. punktowe czy kołowe. Analizując dane będziemy chcieli stworzyć funkcje znajdujące np. pięciu najbogatszych miliarderów z każdego roku, tworząc przy tym od razu wykresy kołowe przedstawiające udział w ogólnym majątku. Analizę będziemy również przeprowadzać według określonej branży czy danego kraju. Będzie się ona również skupiać na poszczególnych latach oraz na ich podstawie przedstawiać interesujące nas informacje.

Etapy tworzenia projektu

Dzień 1

14.05.2022 r. (sobota)

Pracę nad projektem rozpoczęliśmy od znalezienia interesującej nas ramki danych oraz pobrania jej. Ramka danych była dostępna do pobrania w formacie *csv.* Zawierała ona informacje na temat 2600 miliarderów jednak tylko na podstawie 2022 roku.

```
billionaires <- read.csv("Forbes_Billionaires_2022.csv", sep=",", dec=".")
billionaires <- billionaires[,-1]
networth <- billionaires$networth
billionaires <- billionaires[,-3]
billionaires <- cbind(billionaires, "2022"= networth)
```

Aby było wiadomo, iż pobrana ramka zawiera informacje na temat 2022 roku, nazwa kolumny *networth* została zmieniona na *2022.*

Chcąc zgromadzić co najmniej 30000 danych musieliśmy zagłębić się w wybraną tematykę i znaleźć dane na temat innych lat. Podczas poszukiwania ramek z innych lat napotykaliśmy takie, które były również dostępne w formacie *csv* oraz takie, które musieliśmy pobrać za pomocą funkcji służących do pobrania tabeli z internetu, takich jak *html_nodes* i *html_table.* Przykładowo:

```
sciezka <- paste("https://stats.areppim.com/listes/list_billionairesx17xwor.htm")
path <- "/html/body/table"
nodes <- html_nodes(read_html(sciezka), xpath=(path))
bil17 <- html_table(nodes)
billionaires17 <- data.frame(bil17[[1]])
billionaires17 <- billionaires17[-1,]
billionaires17 <- billionaires17[-1,]

names17 <- billionaires17$X2
names17 <- gsub(" ", "", names17)
names17 <- toupper(names17)

networth17 <- 1:2600
for(i in 1:length(names)){
  x <- 0
  for(j in 1:length(names17)){
    if(names[i] == names17[j]){
      networth17[i] <- billionaires17[j,3]
      x <- 1
    }
    if(x == 0){
      networth17[i] <- NA
    }
  }
}
```

Z każdej ramki pobieraliśmy tylko kolumnę z majątkiem z danego roku do wektora, w odpowiedniej kolejności, tak aby później pasował do pierwotnej ramki danych. Niektóre dane o majątku zawarte w innych ramkach miały strukturę różniącą się od tej, która była w ramce z 2022 roku, dlatego musieliśmy ją odpowiednio zmienić. Przeważnie polegało to na tym, że dane w wektorze były numeryczne np. „171.5”, natomiast w ramce dane były formatu ciągu znaków np. „$219 B”. Do zamiany struktury użyliśmy polecenia *stri_paste()*, dodając odpowiednio z przodu „\\$” oraz „ B” z tyłu.

```
networth19 <- stri_paste("$", networth19, sep="")
networth19 <- stri_paste(networth19, "B", sep=" ")

billionaires <- cbind(billionaires, "2019" = networth19)
```

Na koniec, gotową ramkę danych, na podstawie której opiera się cały projekt wyeksportowaliśmy do pliku *csv* do odpowiedniego folderu.

```
write.csv(billionaires, "C:/Users/Piotrek/Desktop/Uczelnia/Programowanie w R/Projekt Końcowy/Forbes_Billionaires_Projekt.csv")
```

Dzień 2.

16.05.2022 r. (poniedziałek)

Tego dnia ustaliliśmy plan działania oraz jakie funkcje chcemy zawrzeć w naszym projekcie. Wiele z tych pomysłów przez cały proces tworzenia projektu udało nam się zrealizować, jednak nie wszystkie plany zostały wykonane, natomiast myślimy, że wyczerpaliśmy odpowiednio temat. Na samym początku stworzyliśmy folder, który będzie służył do przechowywania samego projektu, jak i plików wejściowych, z których korzystamy oraz plików wyjściowych zawierających wyniki przeprowadzonej analizy. W tym dniu zaczęliśmy tworzyć pierwszą funkcję, która ma wyszukiwać najbogatszych miliarderów według średniej na rok. Następnie stworzyliśmy osobną ramkę danych, zawierającą nazwiska i wartości majątków na przestrzeni lat w postaci numerycznej, aby można było wykonywać na tych danych działania.

```
names <- billionaires$name #pobranie nazwisk miliarderów do wektora
#ramka ktora posluzy do stworzenia ramki z wartosciami miliarderó jako wartości numeryczne
kasa <- data.frame("2022" = billionaires$x2022, "2021" = billionaires$x2021, "2020" = billionaires$x2020,
                   "2019" = billionaires$x2019, "2018" = billionaires$x2018, "2017" = billionaires$x2017,
                   "2016" = billionaires$x2016, "2015" = billionaires$x2015 )
for (i in 1:2600){ #petla usuwajaca znaki inne jak liczbowe z ramek wartości majatku miliarderów
  for (j in 1:8){
    kasa[i, j] <- gsub("\\$", "", kasa[i, j])
    kasa[i, j] <- gsub(" B", "", kasa[i, j])
  }
}
#zamiana wartości już bez znaków nie liczbowych na typ numeryczny
kasa$x2022 <- as.numeric(kasa$x2022)
kasa$x2021 <- as.numeric(kasa$x2021)
kasa$x2020 <- as.numeric(kasa$x2020)
kasa$x2019 <- as.numeric(kasa$x2019)
kasa$x2018 <- as.numeric(kasa$x2018)
kasa$x2017 <- as.numeric(kasa$x2017)
kasa$x2016 <- as.numeric(kasa$x2016)
kasa$x2015 <- as.numeric(kasa$x2015)
#swtorzenie ramki z nazwiskami miliarderów i ich majatkami jako numerycznymi liczbami
networth <- data.frame("names" = names, kasa)
```

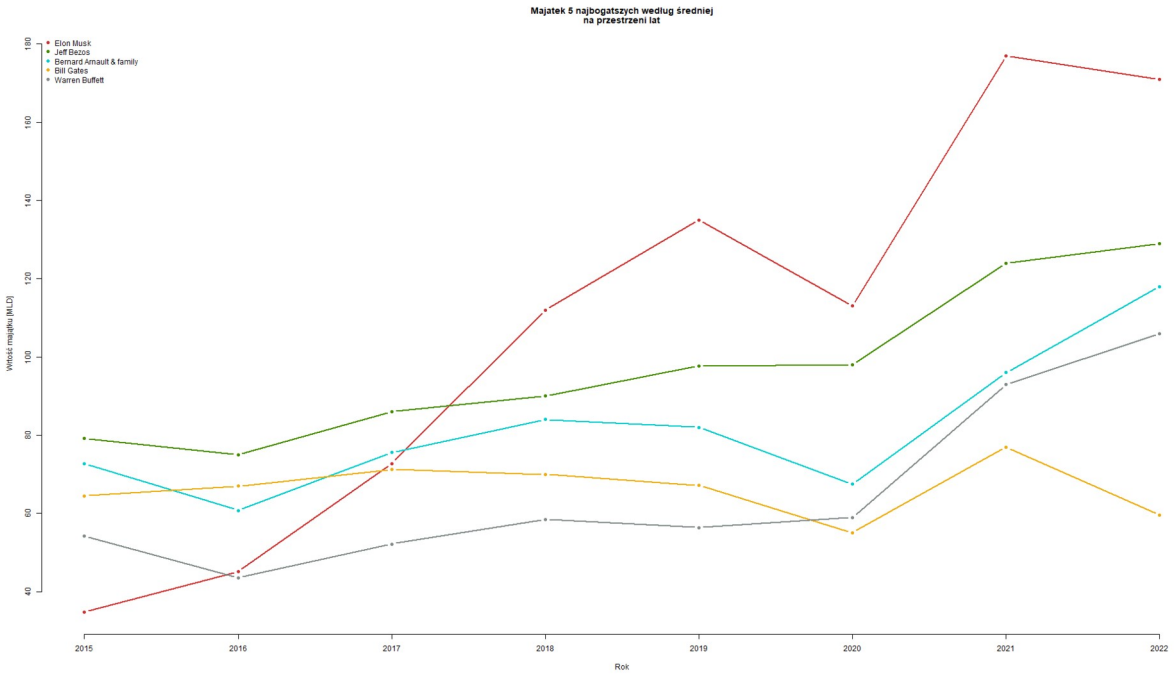
Później za pomocą pętli wyliczaliśmy średnią majątku ze wszystkich lat dla każdego miliardera z osobna i wypisywaliśmy wynik do nowej ramki, w której znajdowało się nazwisko oraz ta wartość średnia, a wszystkie dane posortowaliśmy według wysokości tego majątku, za pomocą funkcji *arrange()*.

```
avg_netw <- data.frame("names" = 1:2600, "AVG" = 1:2600)

#wypełnienie ramki avg_netw
for (i in 1:2600){
  suma <- 0
  for (j in 2:9){
    if(is.na(networth[i, j])){
      suma <- suma
    }
    else{
      suma <- suma + networth[i, j]
    }
  }
  avg_netw[i,1] <- names[i]
  avg_netw[i,2] <- round(suma/8, 2)
}
#posortowanie ramki avg_netw według średniej majątku malejaco
avg_sorted <- arrange(avg_netw, desc(AVG))
```

Kolejnym krokiem było stworzenie nowej ramki danych ze wszystkimi informacjami o miliarderach, zawierającą dodatkowo średnią wartość majątku oraz aktualną jego wartość.

```
AVG_names <- avg_sorted[1:25,1]
AVG_wiek <- 1:25
AVG_kraj <- 1:25
AVG_inc <- 1:25
AVG_anw <- avg_sorted[1:25, 2]
AVG_act <- 1:25
for (i in 1:25){
  for (j in 1:2600){
    if(avg_sorted[i, 1] == billionaires[j, 2]){
      AVG_wiek[i] <- billionaires[j, 3]
      AVG_kraj[i] <- billionaires[j, 4]
      AVG_inc[i] <- billionaires[j, 6]
      AVG_act[i] <- billionaires[j, 7]
    }
  }
}

#sklejenie wszystkich informacji do jednej ramki danych
AVG_anw <- stri_paste("$", AVG_anw, sep="")
AVG_anw <- stri_paste(AVG_anw, "B", sep=" ")
Najbogatsi_wg_avg <- data.frame("Nazwisko" = AVG_names, "Wiek" = AVG_wiek, "Kraj" = AVG_kraj,
                                "Branza" = AVG_inc, "Srednia wartosc na przestrzeni lat" = AVG_anw, "Aktualna wartosc" = AVG_act)
#wysylanie ramki do pliku csv i doc
```

Ramka ta została wysłana do pliku *csv* i do pliku *doc* oraz stworzyliśmy wykres z majątkiem pięciu najbogatszych miliarderów na przestrzeni lat.



Dzień 3.

19.05.2022 r. (czwartek)

Tego dnia zajęliśmy się tworzeniem funkcji wyszukującej progresy i regresy na przestrzeni lat. Zrobiliśmy to za pomocą dwóch zagnieżdżonych pętli, sprawdzając największy progres i regres dla każdego miliardera z osobna, porównując wartości sąsiadujących ze sobą lat. Wyniki zapisywaliśmy do wektorów. Do jednego progres

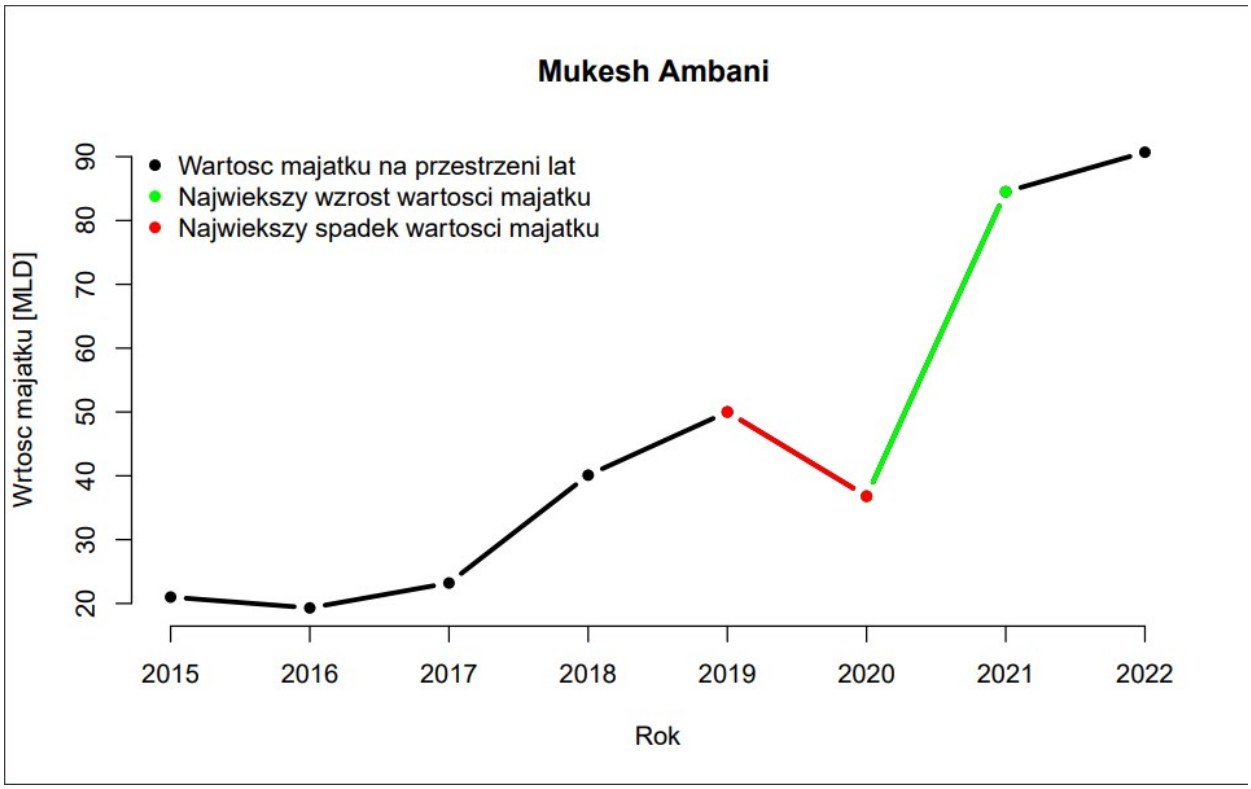
oraz na przestrzeni jakich lat miało to miejsce i do drugiego regres również z odpowiednimi latami.

```
for(i in 1:2600){
  minn <- 300
  maxx <- 0
  a <- 0
  for(j in 9:3){
    if(!is.na(networth[i,j - 1]) & !is.na(networth[i,j])){
      a <- networth[i, j - 1] - networth[i, j]
    }
    else{
      a <- a
    }
    if(a > maxx){
      maxx <- a
      rokpr[i] <- stri_paste(colnames(networth)[j], " - ", colnames(networth)[j - 1])
    }
    else if(a < minn){
      minn <- a
      rokrg[i] <- stri_paste(colnames(networth)[j], " - ", colnames(networth)[j - 1])
    }
  }
  progres[i] <- maxx
  regres[i] <- minn
}
#wyrzucenie znaku  "X" z wektorów lat żeby otrzymać czysty wynik np 2021-2022 a nie X2021-X2022
rokpr <- gsub("X", "", rokpr)
rokrg <- gsub("X", "", rokrg)
```

Następnie stworzyliśmy dwie ramki, jedną łączącą w sobie wektory zawierające dane na temat progresów i drugą gromadzącą dane na temat regresów. Kolejnym krokiem było stworzenie nowej ramki danych, która łączyła w sobie dwie poprzednie ramki oraz dane na temat miliardera, takie jak: nazwisko, wiek oraz branża. Dane w ramce zawierającej wszystkie potrzebne nam informacje zostały posortowane według progresów, od największego do najmniejszego.

```
Progres <- data.frame("Name" = billionaires$name, "MaxProgres" = progres, "Lata progresu" = rokpr)
Regres <- data.frame("Name" = billionaires$name, "MaxRegres" = regres, "Lata regresu" = rokrg)

#stworzenie spójnej ramki danych zawierającej szczegółowe informacje na temat miliarderów i scalającą ich progresy i regresy w jednym miej
zestawieniePr <- data.frame("Nazwisko" = Progres$Name, "Wiek" = billionaires$age[which(billionaires$name == Progres$Name)],
                    "Branza" = billionaires$industry[which(billionaires$name == Progres$Name)], "Najwiekszy progres majatku" = Progres$M
                    "Lata miedzy ktorymi najbardziej zwiekszyl/a majatek" = Progres$Lata.progresu, "Najwiekszy regres majatku" = Regres$
                    "Lata miedzy ktorymi najbardziej zmniejszyl/a majatek" = Regres$Lata.regresu[which(Regres$Name == Progres$Name)])

#ramke posortowaliśmy według największych progresów majatku malejąco i wysłaliśmy ją do pliku
zestawieniePr <- arrange(zestawieniePr, desc(Najwiekszy.progres.majatku))
```

Po odpowiednim przygotowaniu danych stworzyliśmy wykresy majątku na przestrzeni lat, zaznaczające odpowiednio progresy i regresy, dla 10 największych progresów majątku w ciągu kilku lat. Poniżej przykładowy wykres:



Dzień 4.

22.05.2022 r. (niedziela)

Tego dnia zajęliśmy się analizą najbogatszych miliarderów z każdego roku. Na początku stworzyliśmy ramkę danych *Najbogatsi*, w której znalazło się po dziesięć nazwisk, najbogatszych dla każdego roku 2015 – 2022.

```
Najbogatsi <- data.frame("2022"=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0), "2021"=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0),  "2020"=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
                         "2019"=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0), "2018"=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0), "2017"=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
                         "2016"=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0), "2015"=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0))

#dla każdego roku szukamy 10 najbogatszych porównując majatki a następnie usuwając majątek najbogatszej osoby po każdym przejściu petli
#aby jej nie zduplikowało
for(i in 2:9){
  networth_pomoc <- networth
  j <- 1
  while(j <= 10){
    maxx <- 0
    id <- NA
    for(k in 1:2600){
      if(!is.na(networth_pomoc[k, i]) & networth_pomoc[k, i] > maxx){
        maxx <- networth_pomoc[k, i]
        id <- k
      }
      else{
        maxx <- maxx
      }
    }
    Najbogatsi[j, i-1] <- id
    j <- j + 1
    networth_pomoc[id, i] <- NA
  }
}
```

Następnie stworzyliśmy ramkę danych *najbogatsi_rok*, do której w pętli były wypisywane dane dziesięciu najbogatszych miliarderów z zakresu lat 2015 – 2022 (każdy rok osobno przy każdej iteracji pętli). Każda ramka danych była wysyłana do pliku *csv*.

```r
for(i in 1:8){
  for(j in 1:10){
    Najbogatsi_rok[j,] <- billionaires[which(billionaires$rank == Najbogatsi[j, i]), c(1:6, i+6)]
  }
  #następnie w zależności, który rok jest analizowany, ramka jest wysyłana do odpowiedniego pliku, a wykresy zestawienia 10 najbogatszych
  #oraz ich udziału w ogólnym majątku tego roku są wysyłane do pdf
  if(i == 1){
    write.csv(Najbogatsi_rok, "C:/Users/Piotrek/Desktop/Uczelnia/Programowanie w R/Projekt Końcowy/Najbogatsi w roku/Najbogatsi_2022.csv")
```

Kolejnym krokiem było stworzenie wykresów na podstawie zgromadzonych danych. Pierwszym z nich był wykres słupkowy przedstawiający majątek dziesięciu najbogatszych osób dla odpowiedniego roku.

```r
Najbogatsi_rok$Wartosc.majatku.w.tym.roku <- gsub("\\$", "", Najbogatsi_rok$Wartosc.majatku.w.tym.roku)
Najbogatsi_rok$Wartosc.majatku.w.tym.roku <- gsub(" B", "", Najbogatsi_rok$Wartosc.majatku.w.tym.roku)
Najbogatsi_rok$Wartosc.majatku.w.tym.roku <- as.numeric(Najbogatsi_rok$Wartosc.majatku.w.tym.roku)
d <- max(Najbogatsi_rok$Wartosc.majatku.w.tym.roku)
barplot(Najbogatsi_rok$Wartosc.majatku.w.tym.roku, beside = TRUE, las = 2, col = YOR2, ylab = "Wartość majatku [MLD]", ylim = c(0,d +20),
        main = "NAJBOGATSI ROK 2022", names.arg = Najbogatsi_rok$Nazwisko, fon.lab = 3)
```

Aby stworzyć drugi interesujący nas wykres, musieliśmy zsumować majątki wszystkich miliarderów z danego roku oraz majątki tych dziesięciu najbogatszych, a następnie mając te dane stworzyliśmy wykres kołowy, na którym widnieje udział procentowy najbogatszych miliarderów do całej reszty w tym roku.

```r
all <- networth[,2]
suma <- 0
#obliczamy sume ogólna majatków w tym roku aby stworzyć statystyke udziały 10 najbogatszych w danym roku
for(k in 1:2600){
  if(!is.na(all[k])){
    suma <- suma + all[k]
  }
}
suma
#liczymy procentowy udział miliarderów w majatku tego roku
sum <- sum(Najbogatsi_rok$Wartosc.majatku.w.tym.roku)
percentage <- sum/suma*100
percentage <- round(percentage, 2)
percentage <- stri_paste("Ich udział w ogólnym majątku to ", percentage, '%')
#wykres kołowy pokazujący obliczone wyżej dane
pie(c(suma, sum), labels = c("Reszta miliarderów", "10 Najbogatszych w tym roku"), edges = 400, las = 2, col = c("aquamarine1", "azure2"),
    init.angle = 0, main = c("Zestawienie sumy majątków 10 najbogatszych ludzi", "w roku 2022 z resztą miliarderów"),
    xlab = percentage)
```

Powyżej opisane kroki, powtarzają się w pętli dla każdego roku z osobna, a dane wysyłane są do odpowiednich plików, nazwanych *Najbogatsi_XXXX* gdzie w miejscu „*XXXX*” znajduje się rok, z którego pochodzą dane.

Zarówno wykresy kołowe jak i słupkowe z każdego roku zostały wyeksportowane do jednego pliku *PDF*. Wyniki tej analizy znajdują się w folderze *Najbogatsi w roku*.

Dzień 5.

26.05.2022 r. (czwartek)

Tego dnia zajęliśmy się zestawieniem krajów z największą ilością miliarderów. Przy tej funkcji napotkaliśmy pewny problem, a mianowicie było nim wypisanie krajów pojawiających się w ramce danych *billionaires* do wektora jako unikatowe nazwy bez powtórzeń, oraz zliczoną liczbę miliarderów z każdego kraju co robiliśmy niezliczoną ilość razy różnymi pętlami, próbując zoptymalizować działanie tej funkcji. Problem ten udało się rozwiązać, gdy przypomnieliśmy sobie zajęcia z danych typu *factor*. Do zmiennej *country* wpisywaliśmy kraje jako factor z kolumny

billionaires$country, a następnie do wektora *kraj* wpisaliśmy poziomy (*levels*) ze zmiennej factor *country,* co rozwiązało nasz problem z optymalizacją funkcji oraz wypisaniem unikatowych nazw krajów. Dodatkowo zmienną *country* nadpisaliśmy poleceniem *table()* co wypisało nam odpowiadającą krajom, zliczoną liczbę miliarderów stamtąd pochodzących.

```
country <- as.factor(billionaires$country) #pobranie krajów z tabelki
kraj <- levels(country) #pobranie krajów jako unikatowe wartości
country <- table(country) #pobranie liczby miliarderów z każdego kraju
liczba_w_kraju <- as.numeric(country) #przedstawienie tej liczby jako wartość numeryczna
```

Kolejnym krokiem było stworzenie za pomocą pętli ramki danych wypisującą najbogatszego przedstawiciela kraju.

```
najbogatszy <- 1:75
najbogatszy

#tworzymy pomocniczą ramkę z aktualnymi majątkami miliarderów i szukamy najbogatszego w każdym kraju aktualnie
for(j in 1:75){
  ct <- kraj[j]
  x <- data.frame("names" = networth$names, "X2022" = networth$X2022)
  i <- max(x$X2022[which(billionaires$country == ct)])
  najbogatszy[j] <- x$names[which(x$X2022 == i)]
}
```

Następnie w nowej ramce danych znalazły się dane na temat każdego kraju, liczby występujących w nim miliarderów oraz najbogatszego miliardera aktualnie w tym kraju. Ramka ta została posortowana malejąco według ilości miliarderów.

```
#tworzenie ramki danych z informacją o kraju, ile ma miliarderów i który miliarder z tego kraju jest aktualnie najbogatszy
Countries <- data.frame("Kraj" = kraj, "Liczba miliarderow w kraju" = liczba_w_kraju, "Aktualnie najbogatszy w kraju" = najbogatszy)
Countries <- arrange(Countries, desc(Liczba.miliarderow.w.kraju)) #sortujemy ramkę względem liczby miliarderów malejąco
```

Na podstawie zgromadzonych danych chcieliśmy stworzyć wykres kołowy procentowego udziału krajów w ogólnej liczbie miliarderów, jednak krajów było bardzo dużo, przez co wykres był nieczytelny. Zdecydowaliśmy więc, że na wykresie znajdzie się procentowy udział dwunastu krajów z największą ilością miliarderów oraz udział pozostałych krajów podzielonych na grupy zawierające od 1 – 20 miliarderów oraz od 21 – 40 miliarderów, dzięki czemu wykres jest czytelniejszy i możemy z niego bez problemu odczytać dane. Stworzone wykresy zostały wyeksportowane do *pdf* oraz do *png*, natomiast ramka danych do pliku *csv.*

```
kraje1 <- Countries$Liczba.miliarderow.w.kraju[1:12]
kraje2 <- sum(Countries$Liczba.miliarderow.w.kraju[13:23])
kraje3 <- sum(Countries$Liczba.miliarderow.w.kraju[23:75])

#wysyłamy ramkę danych do pliku csv
write.csv(Countries, "C:/Users/Piotrek/Desktop/Uczelnia/Programowanie w R/Projekt Końcowy/Zestawienie krajów/Kraje z największą ilością miliarderów.csv

#sumujemy liczbę miliarderów z niektórych krajów
suma_20_40 <- sum(Countries$Liczba.miliarderow.w.kraju[13:23])
suma_1_20 <- sum(Countries$Liczba.miliarderow.w.kraju[24:75])

#liczmy procentowy udział pierwszych 12 krajów oraz reszty w liczbie miliarderów aby utworzyć czytelny wykres
piepercent <- c(round(Countries$Liczba.miliarderow.w.kraju[1]/2600*100, 1), round(Countries$Liczba.miliarderow.w.kraju[2]/2600*100, 1), round(Countries
                round(Countries$Liczba.miliarderow.w.kraju[4]/2600*100, 1), round(Countries$Liczba.miliarderow.w.kraju[5]/2600*100, 1), round(Countries
                round(Countries$Liczba.miliarderow.w.kraju[7]/2600*100, 1), round(Countries$Liczba.miliarderow.w.kraju[8]/2600*100, 1), round(Countries
                round(Countries$Liczba.miliarderow.w.kraju[10]/2600*100, 1), round(Countries$Liczba.miliarderow.w.kraju[11]/2600*100, 1), round(Countri
                round(suma_20_40/2600 * 100, 1), round(suma_1_20/2600*100, 1))

#wybieramy paletę kolorów
paleta <- choose_palette()
view(paleta)
pa <- paleta(14)

#tworzymy wykres z zestawionymi zebranymi informacjami i wysyłamy go do pdfa oraz png
pdf("C:/Users/Piotrek/Desktop/Uczelnia/Programowanie w R/Projekt Końcowy/Zestawienie krajów/Zestawienie krajów.pdf", height = 8, width = 14)
pie(c(kraje1, kraje2, kraje3), labels = stri_paste(piepercent, "%"), edges = 600, col = pa, main = c("Zestawienie krajów wedlug liczby", "miliarderów p
legend("bottomleft", c(Countries$Kraj[1:12], "Kraje z l. miliarderów między 20 - 40", "Kraje z l. miliraderów między 1 - 20"),
       col = pa, bty = "n", pch = 15)
dev.off()
```

Dzień 6.

29.05.2022 r. (niedziela)

Dzisiejszego dnia stworzyliśmy funkcję, która zwraca dane konkretnej wyszukanej osoby. Funkcja ta zbiera informacje na temat osoby, która została podana na jej wejście (np. Elon Musk), a następnie wypisuje je kolejno za pomocą funkcji printf(). Funkcja jest o tyle ciekawa, że postanowiliśmy w funkcji zawrzeć przeliczanie majątku miliardera z dolarów na polską walutę PLN, po aktualnym kursie dolara. W funkcji pobierany jest z internetu aktualny kurs dolara przy pomocy biblioteki rvest oraz czas, w którym ten kurs jest pobierany, przy pomocy funkcji Sys.time().

```
#pobieramy z internetu aktualny kurs dolara
sciezka <- paste("https://www.bankier.pl/waluty")
path <- "/html/body/div[3]/div[1]/div[2]/div[1]/div[1]/div[2]/div/div[4]/div[2]/table"
#sciezka <- paste("https://internetowykantor.pl/kurs-dolara/")
#path <- "/html/body/div[1]/div[1]/div[2]/div[3]/div[1]/div[2]/div[1]/span[2]"
nodes <- html_nodes(read_html(sciezka), xpath=(path))
kurs <- html_table(nodes)
kurs <- kurs[[1]]
#po pobraniu kursu pobieramy czas pobrania kursu
czas <- Sys.time()
#pobrana tabele zmieniamy w ramke danych
kurs <- as.data.frame(kurs)
#wybieramy z niej kurs dolara
dolar <- as.vector(kurs[1,2])
#zamieniamy przecinek na kropke a nastepnie sam dolar na wartość numeryczna aby można było przeliczyć majatek na PLN
dolar <- gsub("\\,", '\\.', dolar)
dolar <- as.numeric(dolar)
```

Funkcja przed wypisaniem informacji przelicza majątek na PLN i wyświetla to razem z resztą informacji. Ponadto funkcja tworzy skrót od kraju pochodzenia wyszukanej osoby, przy pomocy biblioteki stringi, sprawdza ona przy tym czy nazwa kraju jest jedno-, dwu- czy trzy-członowa i tworzy odpowiedni to tego skrót. Do tworzenia owej funkcji bardzo pomocne okazały się pliki z zadań projektowych nr 6, które były do samodzielnego zrealizowania w domu. Ostatecznie funkcji zgromadzone i przetworzone dane wypisuje za pomocą polecenia printf(), czasem wspomaganego poleceniem stri_paste(). Ponadto, w niektórych danych zostały zastosowane formaty w jakich te dane mają być wypisane (konkretnie chodzi o czas pobrania kursu dolara).

```
print(sprintf("Aktualna wartość majątku: %s", p_worth))
print(paste(sprintf("Aktualna wartość w PLN: %s", networth_PLN), sprintf("MLD zł")))
print(sprintf("Aktualny kurs dolara: %g", dolar))
print(sprintf("Data pobrania kursu: %s", format(czas, format = "%d %B %Y %H:%M")))
print(sprintf("Główne źródło dochodu: %s", zrodelko))
print(paste(sprintf("Największy progres majątku na przestrzeni lat: %s", "$"), sprintf("%s", p_progres), sprintf("%s", "B")))
```

W funkcji oprócz wypisania danych, tworzony jest wykres majątku wyszukanej osoby, na przestrzeni lat 2015 – 2022, wraz z zaznaczonym największym spadkiem oraz skokiem majątku pomiędzy latami.

Dzień 7.

02.06.2022 r. (czwartek)

Dzisiejszego dnia zajęliśmy się funkcją nazwaną *powyzej_ponizej* podanego roku życia. Do tej funkcji należy podać dwa argumenty, pierwszym z nich jest wiek względem którego funkcja ma przeprowadzić analizę, drugim argumentem jest słowo *powyzej* lub *ponizej*. Aby funkcja działała bez zarzutu jest ona uodporniona na błędne wpisanie słów *powyzej*, *ponizej*, tzn. na odstępy między literami czy różną wielkość liter w słowie np. *Po wyZ ej*.

```
powyzej_ponizej <- function(l, po){
  w <- l
  #usuwamy dodatkowe spacje z wpisanego słowa oraz zmieniamy litery jako wielkie
  po <- gsub(" ", "", po)
  po <- gsub("  ", "", po)
  po <- gsub("   ", "", po)
  po <- toupper(po)
```

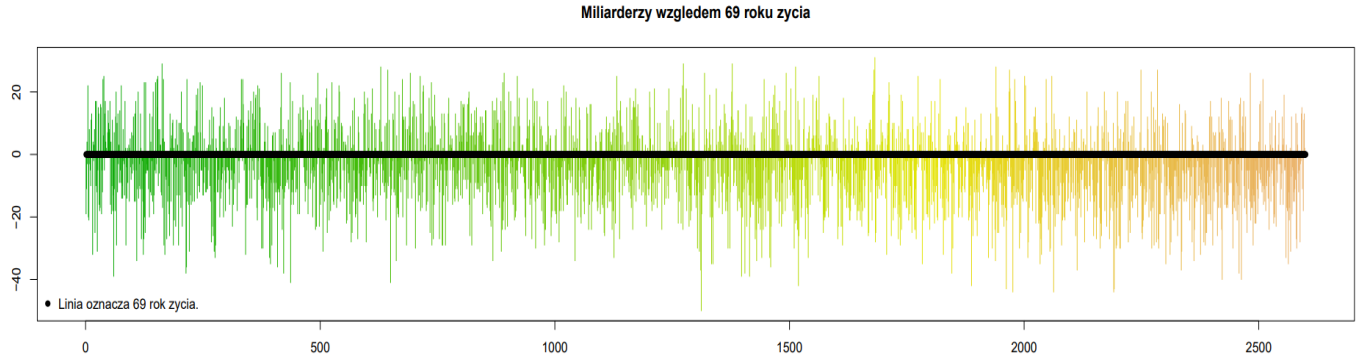
W głównej funkcji tworzymy kolejną funkcje potrzebną nam do narysowania histogramu, a mianowicie ta funkcja wypisuje nam wiek miliardera po danym numerem ID, co jest wykorzystane w histogramie, gdzie oś X to zbiór od 1-2600, a oś Y to wyświetlana funkcja, która jako argument przyjmuje wartości ze zbioru X. Dodatkowo w naszej funkcji wiek miliarderów jest zaniżony o wiek, który się wybrało żeby histogram przedstawił porównanie do wybranego wieku a nie do 0 na osi.

```
f <- function(x){
  as.numeric(billionaires$age[x] - l) #wiek pomniejszyliśmy o wyszukany aby ładnie wyglądało to na histogramie
}
f(2) #sprawdzenie działania funkcji
x <- 1:2600 #wektor pomocniczy do wykresu

#tworzymy histogram wieku miliarderów względem wybranego wieku i wysyłami go do pdfa
pdf("C:/Users/Piotrek/Desktop/Uczelnia/Programowanie w R/Projekt Końcowy/Rok życia/Histogram według wieku.pdf", width = 20, height = 5)
plot(x, f(x), type = "h", col = terrain.colors(3600), xlab="", ylab="", main = stri_paste("Miliarderzy względem ", l, " roku życia"))
lines(x, rep(0, 2600), type = "b", pch = 16, col = "black", lwd = 1)
legend("bottomleft", stri_paste("Linia oznacza ", l, " rok życia."), col = "black", bty = "n", pch = 16)
dev.off()
```

Następnie w zależności od wybranego słowa *ponizej* lub *powyzej* za pomocą pętli tworzone są trzy osobne wektory, które zawierają kolejno: dane na temat nazwiska miliardera poniżej lub powyżej wybranego wieku, jego aktualne miejsce w rankingu oraz wiek. Ze stworzonych wektorów usuwane są wartości *NA* i na ich podstawie tworzona jest ramka danych odpowiednio nazwana *POWYZEJ* lub *PONIZEJ*, do której jako pierwszy wiersz dodawany jest komunikat *„wybrany wiek: „* oraz wiek podany do funkcji, aby było wiadomo otwierając ramkę *csv* jaki wiek został wybrany przy wywoływaniu funkcji.

```
#dodajemy do ramki do pierwszego wiersza info na temat jaki wiek był wybrany do wyszukania osób powyżej tego roku życia
w_rok <- c("Wybrany wiek:", w, NA)
POWYZEJ <- rbind(w_rok, POWYZEJ)
#wysyłamy ramkę do pliku
write.csv(POWYZEJ, "C:/Users/Piotrek/Desktop/Uczelnia/Programowanie w R/Projekt Końcowy/Rok życia/Miliarderzy powyzej wybranego roku zycia.csv")
```

Przykładowy histogram stworzony po wywołaniu funkcji względem wieku



Dzień 8.

05.06.2022 r.

W dniu dzisiejszym, zrobiliśmy ostatnią już funkcję w naszym programie, która przeprowadza analizę wyszukanej branży. Funkcja podobnie jak funkcja analizująca miliarderów względem wieku, jest odporna na wpisanie nazwy branży z błędami w stylu, różna wielkość liter, czy odstępy między nimi. Jako argument owej funkcji należy podać nazwę branży, która nas interesuje (np. „TECHNOLOGY”). Funkcja ta tworzy ramkę danych branża, w której wypisuje dane na temat każdego miliardera, pracującego czy też specjalizującego się w wyszukanej branży. Ramka ta zostaje następnie wysłana do odpowiednio podpisanego pliku *csv*.

```
#pobieramy do pomocniczej ramki ramke billionaires żeby zamienić w niej nazwy branzy tak aby funkcja mogła wyszukać tą która nas interesuje
bill <- billionaires
bill$industry <- gsub(" ", "", bill$industry)
bill$industry <- toupper(bill$industry)
bill
#pobieramy dane na temat miliarderów z danej branży i ich aktualnego majatku
branza <- bill[which(bill$industry == b), c(1:5, 7)]
colnames(branza) <- c("Aktualny ranking", "Nazwisko", "Wiek", "Kraj pochodzenia", "Zrodlo dochodu", "Aktualny.majatek") #zamieniamy nazwy kolumn
branza <- rbind(c(NA, stri_paste("Wybrana branza: ", b), NA, NA, NA, NA), branza) #dodajemy pierwszy wiersz który bedzie miał informacje na tema

#wysylamy stworzoną ramkę do pliku
write.csv(branza, "C:/Users/Piotrek/Desktop/Uczelnia/Programowanie w R/Projekt Końcowy/Analiza branzy/Miliarderzy z wybranej branzy.csv")
```

W funkcji następnie tworzona jest ramka danych wartości, która zawiera dane na temat nazwisk i aktualnych majątków miliarderów tym razem w postaci numerycznej, z wybranej branży. Jest tak ponieważ owa ramka danych przyda nam się do obliczenia procentowego udziału danej branży w ogólnym majątku w roku 2022 (aktualne dane). W pętli sumowane są majątki z 2022 roku, miliarderów z każdej branży. Dla każdej branży tworzona jest osobna suma.

```
suma <- data.frame("Branza" = branze, "suma" = rep(0, 18))

#szukamy majatków z danej branży i sumujemy je wpisując do ramki przy odpowiedniej branży
for(i in 1:18){
  for(j in 1:2600){
    if(networth_branza$branza[j] == branze[i]){
      suma[i,2] <- suma[i,2] + networth_branza$`2022`[j]
    }
  }
}
```

Następnie obliczony został procentowy udział wybranej branży w ogólnym zsumowanym majątku z 2022 roku, oraz procentowy udział sumy reszty branż, co zostało przedstawione na wykresie kołowym. Funkcja dodatkowo tworzy przy pomocy polecenia *par(mfrow = c(2,3))*, wykres majątków na przestrzeni lat, 6 najbogatszych miliarderów w tej branży. Ostatecznie zarówno 6 wykresów majątków miliarderów jak i wykres kołowy obrazujący procentowy udział wybranej branży w ogólnym majątku z aktualnego roku, są wysyłane do pliku PDF.

```
suma$Branza <- toupper(suma$Branza)
suma$Branza <- gsub(" ", "", suma$Branza)
suma_wszystkich <- sum(suma[,2]) - suma[which(suma$Branza == b), 2] #sumujemy majątek wszystkich branży oprócz wybranej
#obliczamy procentowy udział naszej branży i reszty branż razem
piepercentage <- c(round(suma_wszystkich/sum(suma[,2]) * 100, 2), round(suma[which(suma$Branza == b), 2]/sum(suma[,2]) * 100, 2))

year <- c(2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022) #wektor pomocniczy do wykresu
pdf("C:/Users/Piotrek/Desktop/Uczelnia/Programowanie w R/Projekt Końcowy/Analiza branzy/Analiza.pdf", height = 20, width = 20)
#tworzymy 6 wykresów na raz, majątków 6 najbogatszych z wybranej branży
par(mfrow=c(2,3), col = "black")
for(i in 1:6){
  plot(year, networth[which(networth$names == wartosci$Nazwisko[i]), c(9:2)], type = "b", col = terrain.colors(20)[15], pch = 16, xlab = "Rok", ylab
       font.main = 2, las = 0, lty = 1, lwd = 3, main = wartosci$Nazwisko[i])
}

#ustawiamy wykresy znowy na wyświetlanie jednego na raz i tworzymy wykres kołowo procentowego udziału wybranej branżu w ogólnej liczbie majątków
par(mfrow = c(1,1), col = "black")
pie(c(suma_wszystkich, suma[which(suma$Branza == b), 2]), labels = stri_paste(piepercentage, "%"), edges = 600, col = c("darkolivegreen", "darkgray")
legend("bottomright", c("Reszta branzy", b), col = heat.colors(2), bty = "n", pch = 15)
dev.off()
```

Dzień 9.

07.06.2022 r.

Dzisiejszego dnia został wykonany plakat promujący nasz projekt. Nie rozumieliśmy dokładnie co rozumieć przez dokument/plakat więc postanowiliśmy zrobić to co umiemy najlepiej. Plakat został wykonany przy pomocy aplikacji Adobe Photoshop 2022 i jest zamieszczony wraz zresztą plików w folderze projektu. Ponadto znajduje się tam folder Plakat, w którym zostawiliśmy składowe, z których plakat został stworzony.

Wnioski

Analiza danych okazuje się być często żmudnym zajęciem, wymagającym godzin kombinowania i myślenia, szczególnie nad kodem jeśli analizę wykonujemy w programie jak RStudio w języku programowania R. Zajęcie to staje się wtedy o tyle trudniejsze, że często niektóre rozwiązania w kodzie okazują się nie działać, być bardzo oporne bądź po prostu nieoptymalne dla bardziej złożonej analizy. Dokłada to kolejne godziny żmudnej i opornej pracy. Jednakże jest to bardzo wdzięczna praca, bowiem jest bardzo intrygująca sama w sobie, zmuszająca do logicznego myślenia ale

przede wszystkim bardzo satysfakcjonująca. Satysfakcja jest o tyle większa im większy problem sprawiło nam jakieś zagadnienie w trakcie analizy i więcej czasu oraz wysiłku nam zabrało. Odnośnie samego projektu uważamy, że najtrudniejszym zadaniem nie była sama analiza (która swoją drogą okazała się bardzo ciekawa), lecz wymyślenie tematu, a następnie przedstawienie go w ciekawy sposób oddający jednocześnie nasze umiejętności w środowisku języka R. Innymi słowy najgorzej było złapać wenę, aby wymyślić zagadnienia, które nie będą ani zbyt „górnolotne" czyli ciekawe lecz zbyt trudne do wykonania ani zbyt „dolnolotne" czyli może i ciekawe lecz nie pokazujące w pełni umiejętności. Ostatecznie jesteśmy zadowoleni z finalnego efektu wykonania projektu i mamy nadzieję, że równie dobrze zostanie odebrany.

Źródła:

- https://www.kaggle.com/datasets/jjdaguirre/forbes-billionaires-2022
- https://stats.areppim.com/listes/list_billionairesx19xwor.htm
- https://stats.areppim.com/listes/list_billionairesx17xwor.htm
- https://stats.areppim.com/listes/list_billionairesx16xwor.htm
- https://stats.areppim.com/listes/list_billionairesx15xwor.htm
- https://www.kaggle.com/datasets/soubenz/forbes-top-billionaires-list-2018
- https://www.kaggle.com/datasets/roysouravcu/forbes-billionaires-of-2021