Paulina Szymanek, 400734, Inżynieria i analiza danych

Sprawozdanie

W projekcie zostały użyte pakiety:

```
#Pakiety
library(tidyverse)
library(ggplot2)
library(graphics)
library(lattice)
library(plotly)
library(rbokeh)
library(gridExtra)
library(viridis)
library(crayon)
library(latticeExtra)
```

1. Wczytywanie i szczegóły danych

```
#Wczytywanie i zapisywanie danych
dane_surowe<-read.csv("marathon.csv", skip=1, header=FALSE, col.names=c("Order",
"Age", "Gender", "Time"))
write.csv(dane_surowe, "Szymanek_dane_surowe.csv")
```

Plik wczytałam z pominięciem nazw kolumn ze względu na nawiasy zawarte w jednej z nazw, które utrudniały samo wczytanie danych oraz późniejszą pracę z nimi.

```
any(is.na(dane_surowe))
dane_surowe
str(dane_surowe)
summary(dane_surowe)
glimpse(dane_surowe)
class(dane_surowe)
typeof(dane_surowe)
names(dane_surowe)
levels(dane_surowe)
dim(dane_surowe)
head(dane_surowe)
tail(dane_surowe)
```

W danych nie ma braków, więc nie było też wymagane usunięcie ich. Wyświetliłam całość danych, ich podsumowanie, typ, wymiary oraz ich początek i koniec.

```
marathon<-dane_surowe%>% arrange(Order)%>%
mutate(Gender=ifelse(as.character(Gender)==" M", "Male", "Female"))
write.csv(marathon, "Szymanek_dane_przeksztalcone.csv")
```

Jako przetworzone dane zapisałam surowe dane z posortowaną rosnąco kolumną „Order” oraz ze zmienioną zawartością kolumny „Gender”, aby ułatwić dalszą pracę z danymi.

2. Praca z danymi

```
#Praca z danymi
salm<-make_style("indianred1")
cat(salm$underline$bold("All data summary:\n"))
summary(marathon)

m_data<-marathon%>% filter(Gender=="Male")
f_data<-marathon%>% filter(Gender=="Female")

cat(salm$underline$bold("Summary for males:\n"))
summary(m_data)
cat(salm$underline$bold("Summary for females:\n"))
summary(f_data)
```

Podsumowanie danych wykonałam dla całości danych oraz dla ich podziałów według płci.

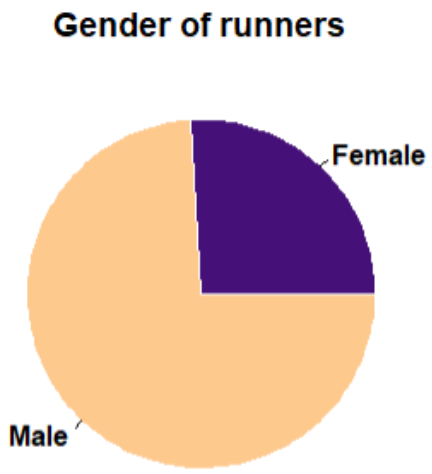
```
marathon%>% group_by(Age)%>% summarize(medianTime=median(Time), meanTime=mean(Time))
marathon%>%group_by(Age)%>% summarize(NumOfRunners=n_distinct(Time))%>%
arrange(desc(NumOfRunners))
marathon%>%group_by(Age)%>% summarize(maxTime=max(Time), minTime=min(Time))%>%
filter(maxTime!=minTime)

co<-viridis_pal(option="magma", direction=-1)(10)
colo<-c(co[8], co[2], co[5])
```

Kolejno wyświetliłam wiek wraz z medianą oraz średnim czasem dla danej wartości. Następnie zliczyłam ilość unikalnych wartości dla wieku i posortowałam je malejąco. Na koniec Wyświetliłam maksymalną i minimalną wartość czasu dla wieków w których te wartości są różne (co oznacza, że dany wiek występuje więcej niż raz.

3. Wizualizacja danych

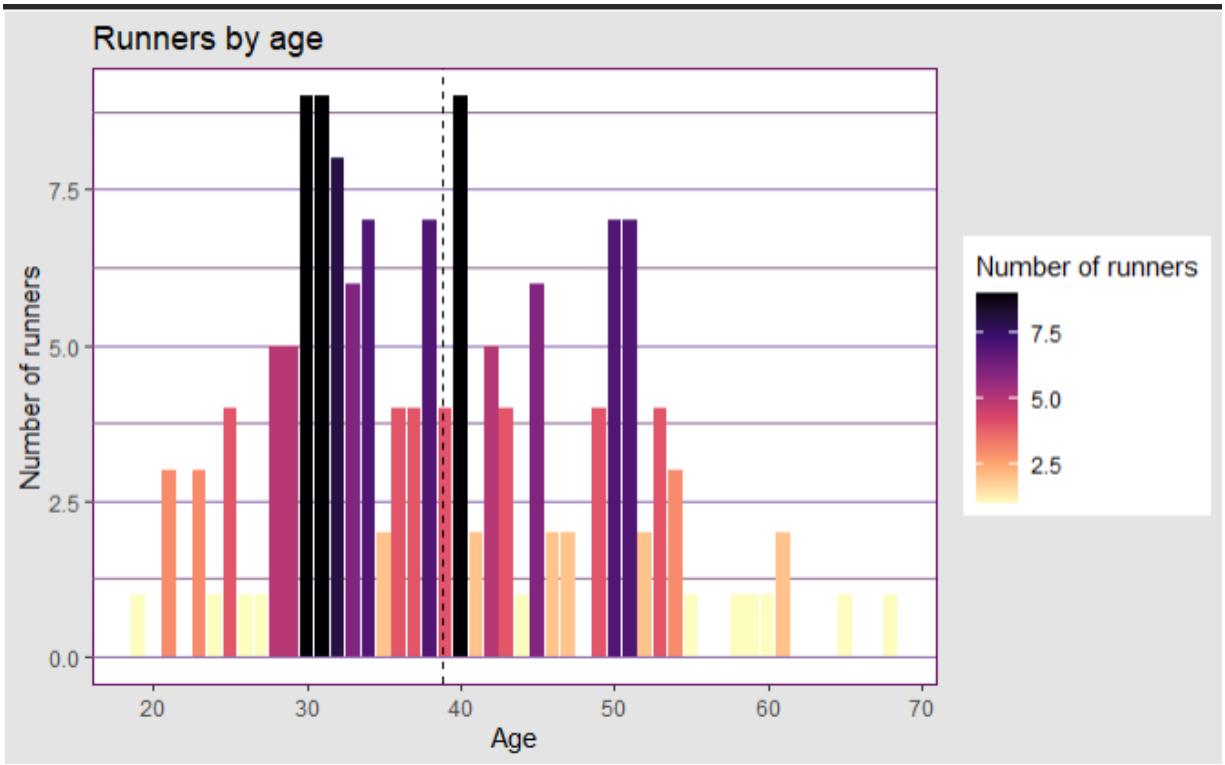
a) Pierwszy wykres został wykonany z użyciem pakietu „graphics”. Jest to wykres kołowy przedstawiający płeć uczestników maratonu. Z wykresu widać, że prawie trzy czwarte uczestników było płci męskiej.

```
#wykres 1
pie(table(marathon$Gender), col=colo, border="white", main="Gender of runners",
font=2)
```



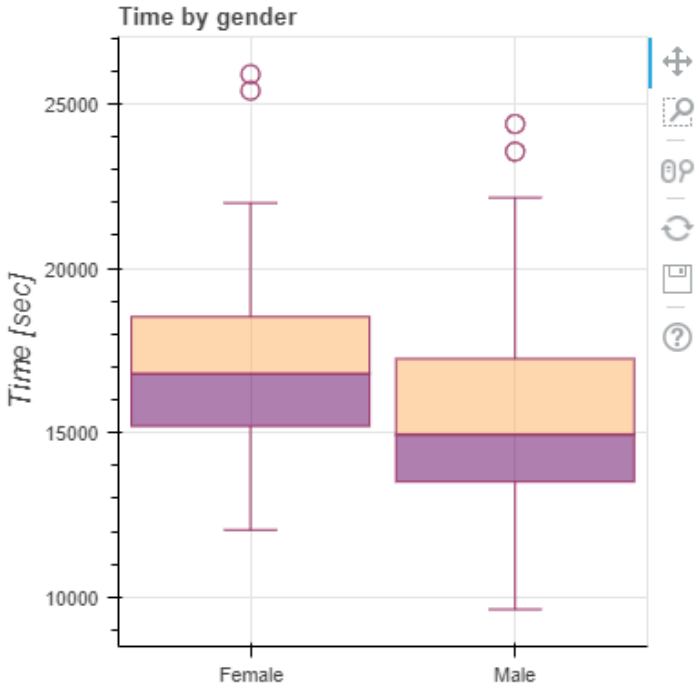
b) Drugi wykres został wykonany z użyciem pakietu „ggplot2”. Jest to wykres kolumnowy przestawiający rozkład uczestników według ich wieku. Najwyższe kolumny widać dla 30, 40 i 50 lat oraz w ich przybliżeniach.

```
#wykres 2
ggplot(marathon, aes(x=Age), y=as.numeric(stat(count)))+
geom_bar(aes(fill=as.numeric(stat(count))))+
 scale_fill_viridis_c(option="magma", direction=-1, name="Number of runners")+
 geom_vline(xintercept=mean(marathon$Age), linetype="dashed")+
 labs(x="Age", y="Number of runners")+ ggtitle("Runners by age")+
 theme(panel.grid.major.x=element_blank(), panel.grid.minor.x=element_blank(),
panel.grid.major.y=element_line(colour="#875faf"),
panel.grid.minor.y=element_line(colour="#875f87"),
panel.background=element_rect(colour="#5f005f", fill="white"),
plot.background=element_rect(fill="#e4e4e4"))
```



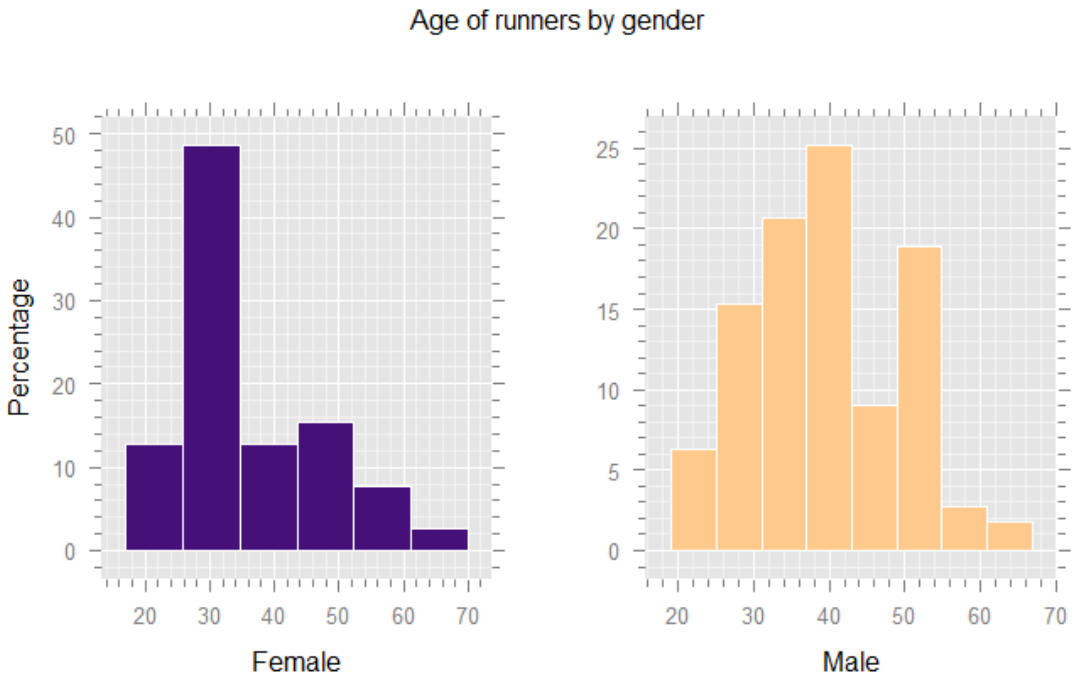
c) Trzeci wykres został wykonany z użyciem pakietu „rbokeh”. Jest to wykres pudełkowy przestawiający rozrzut czasu uzyskanego przez uczestników według ich płci. Wyniki dla uczestników płci męskiej mają mniejszą medianę oraz większy rozrzut pomiędzy kwartylami.

```
#wykres 3
figure(title="Time by gender")%>%
ly_boxplot(x=Gender, y=Time, data=marathon, fill_color=c("#5f005f", "#ffaf5f"),
ylab="Time [sec]", xlab='', line_color="#993366")
```



d) Czwarty wykres został wykonany z użyciem pakietu „lattice”. Jest to histogram przestawiający procent uczestników według ich wieku i płci. Dla uczestników płci żeńskiej niemal połowa była w wieku około 30 lat. Reszta przedziałów wiekowych pozostaje na podobnym poziomie. Dla uczestników płci męskiej widzimy większy rozrzut danych, gdzie największa grupa wiekowa, około 40 lat, stanowi ćwierć uczestników tej płci.

```
#wykres 4
hist1<-histogram(~f_data$Age,group=Gender, data=marathon, xlab="Female",
ylab="Percentage", col=colo[1], border="white",
par.settings=ggplot2like(),lattice.options=ggplot2like.opts())
hist2<-histogram(~m_data$Age,group=Gender, data=marathon, xlab="Male", ylab="",
col=colo[2], border="white",
par.settings=ggplot2like(),lattice.options=ggplot2like.opts())
grid.arrange(hist1, hist2, ncol=2, top="Age of runners by gender")
```



e) Piąty wykres został wykonany z użyciem pakietu „plotly”. Jest to zbiór wykresów liniowych pokazujących zależność średniego czasu uzyskanego przez uczestników i ich wieku. Wykonałam wykresy dla wszystkich uczestników oraz oddzielne według płci. Z wykresu wynika, że najlepsze wyniki uzyskiwali uczestnicy w średnik wieku (35-45 lat). Uczestnicy płci męskiej mieli średnio lepsze wyniki, a uczestnicy płci żeńskiej gorsze wyniki.

```
#wykres 5
time_mean<-marathon%>% group_by(Age)%>% summarize(meanTime=mean(Time))
co<-viridis_pal(option="magma", direction=-1)(10)
time_meanf<-f_data%>% group_by(Age)%>% summarize(meanTimef=mean(Time))
time_meanm<-m_data%>% group_by(Age)%>% summarize(meanTimem=mean(Time))

time_meanf<-merge(time_mean,time_meanf, all=TRUE)%>%select(Age,meanTimef)
time_meanm<-merge(time_mean,time_meanm, all=TRUE)%>%select(Age,meanTimem)

marathon%>% plot_ly(x=~time_mean$Age)%>% layout(title="Time by age",
xaxis=list(title="Age", gridcolor="#d7afaf"), yaxis=list(title="Mean time",
gridcolor="#d7afaf"), legend = list(x=0.1, y=0.9), paper_bgcolor="#e4e4e4")%>%
  add_trace(y=~time_mean$meanTime,mode='lines', name="All",
line=list(color=colo[3]))%>%
  add_trace(y=~time_meanf$meanTimef,mode='lines', name="Females",
line=list(color=colo[1], dash = 'dash'))%>%
  add_trace(y=~time_meanm$meanTimem,mode='lines', name="Males",
line=list(color=colo[2], dash = 'dash'))
```