Projekt 2 Grupowanie

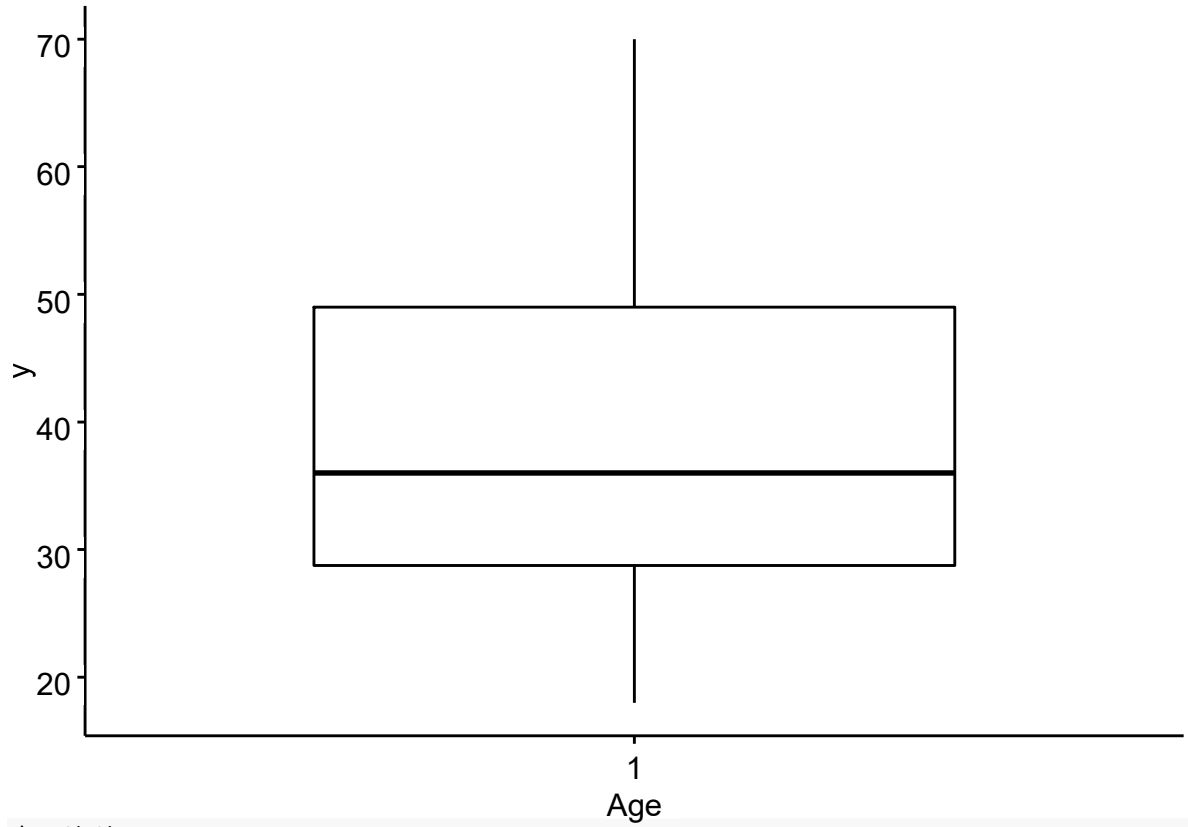
Konrad Komor

```
dane <-read.csv("C:/R/pliki/Mall_Customers.csv",stringsAsFactors = TRUE) dane1<-dane[,-
c(2)] summary(dane)
```

```
##        CustomerID        Gender       Age        Annual.Income..k..
## Min.   : 1.00   Female:112   Min.   :18.00   Min.   : 15.00
## 1st Qu.: 50.75   Male : 88   1st Qu.:28.75   1st Qu.: 41.50
## Median :100.50               Median :36.00   Median : 61.50
## Mean   :100.50               Mean   :38.85   Mean   : 60.56
## 3rd Qu.:150.25               3rd Qu.:49.00   3rd Qu.: 78.00
## Max.   :200.00               Max.   :70.00   Max.   :137.00
## Spending.Score..1.100.
## Min.   : 1.00
## 1st Qu.:34.75
## Median :50.00
## Mean   :50.20
## 3rd Qu.:73.00
## Max.   :99.00
```

```
#Jak widzimy nasze dane zawierają 4 zmienne numeryczne: Annual.Income..k,
#Spending.Score..1.100, Age oraz CustomerID
#Posiadamy też jedną zmienną binarną Gender
#Zmienna Gender zawiera 112 odpowiedzi Female i 88 odpowiedzi Male #CustomerID to Id każdego z
klientów dlatego nie będzie podlegała żadnym analizom
```

```
ggboxplot(dane$Age,xlab="Age")
```




```
dane%>%
summarise(meanAge=mean(Age),
standardDeviationAge=sd(Age))
```

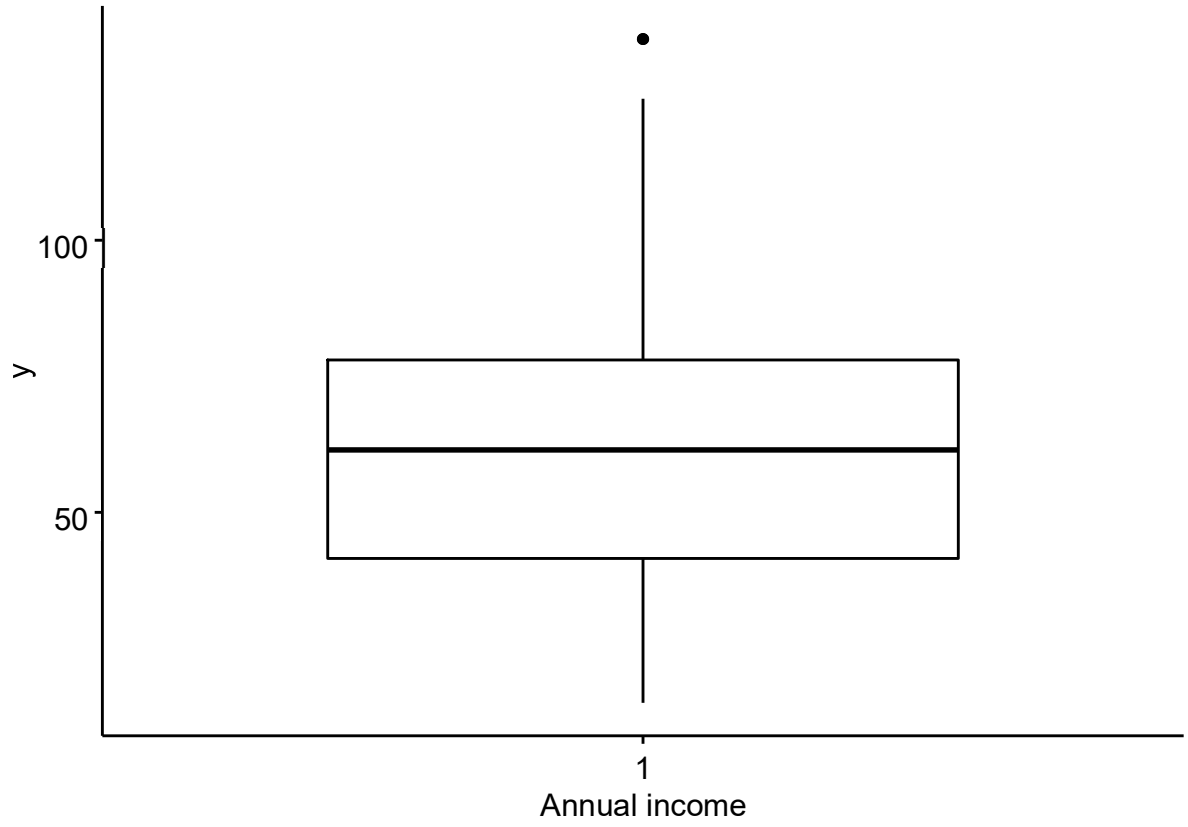
```
##         meanAge standardDeviationAge
## 1         38.85                13.96901
```

```
quantineAge<-quantile(dane$Age) quantineAge
```

```
##       0%     25%     50%     75% 100%
## 18.00 28.75 36.00 49.00 70.00
```

#Średnia wartość zmiennej wiek wynosi 38,85 nie ma wartości odstających. Wiek #odchyla się średnio o 13,96 od średniej.
#Minimalna wartość zmiennej wiek wynosi 18 a maksymalna 70, 50% wszystkich #wartości zmiennej wiek jest równa lub mniejsza od 36

```
ggboxplot(dane$Annual.Income..k.. ,xlab="Annual income")
```




```
dane%>% summarise(meanIncome=mean(Annual.Income..k..),standardDeviationIncome=sd(Annual.Income..k..)) ##
 meanIncome standardDeviationIncome
## 1              60.56                          26.26472
```

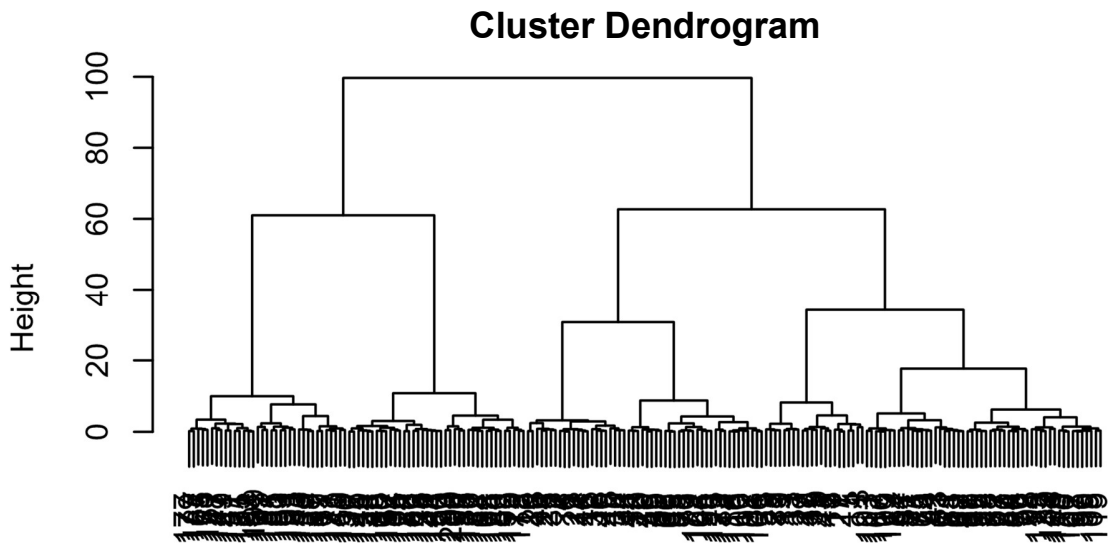
```
quantileIncome<-quantile(dane$Annual.Income..k..) quantileIncome
```

```
##       0%     25%     50%      75% 100%
## 15.0 41.5 61.5 78.0 137.0
```

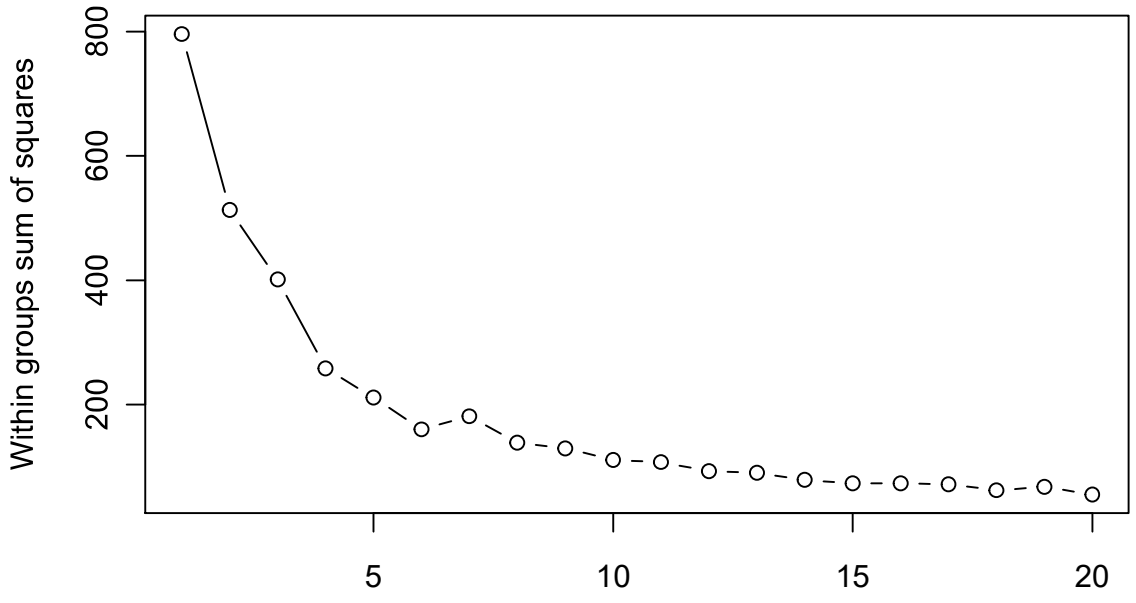
```
#Średnia wartość zmiennej Annual Income wynosi 60,56 jest wartość odstająca. #Annual Income odchyla
się średnio od średniej o 26,26
```

```
dane_std <- dane %>% dplyr::select_if(., is.numeric) %>% scale(.) %>% as.data.frame #dokonuję standaryzacji danych
#Tworzę dendrogram grupowania hierarhicznego z dystansem euklideosowym. Z
#poniższego wykresu można wnioskować, że liczba grup która powinienem wybrać to
#prawdopodobnie 4 lub 6
 d <- dist(dane_std, method = "euclidean") fit <-
hclust(d, method="ward")
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
plot(fit)
```

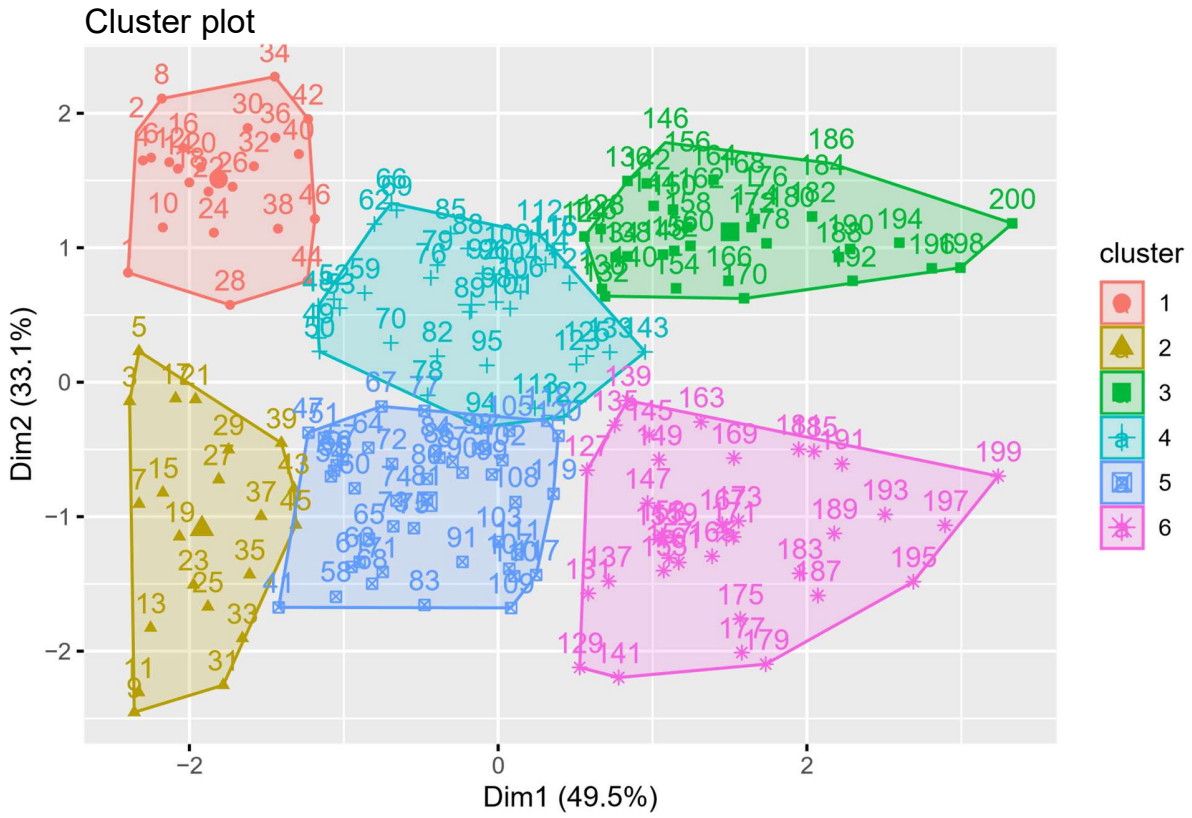



```
#Żeby upewnić się, że wybrałem poprawną liczbę grup tworzę wykres Osypiska
#Jak widzimy 4 grupy nie są najlepszym wyborem. Wykres wypłaszcza się dopiero #powyżej 6 grupy
dlatego też decyduję się na użycie 6 grup wss <- (nrow(dane_std)-1)*sum(apply(dane_std,2,var)) for (i in
2:20) wss[i] <- sum(kmeans(dane_std, centers=i)$withinss)
plot(1:20, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares")
```

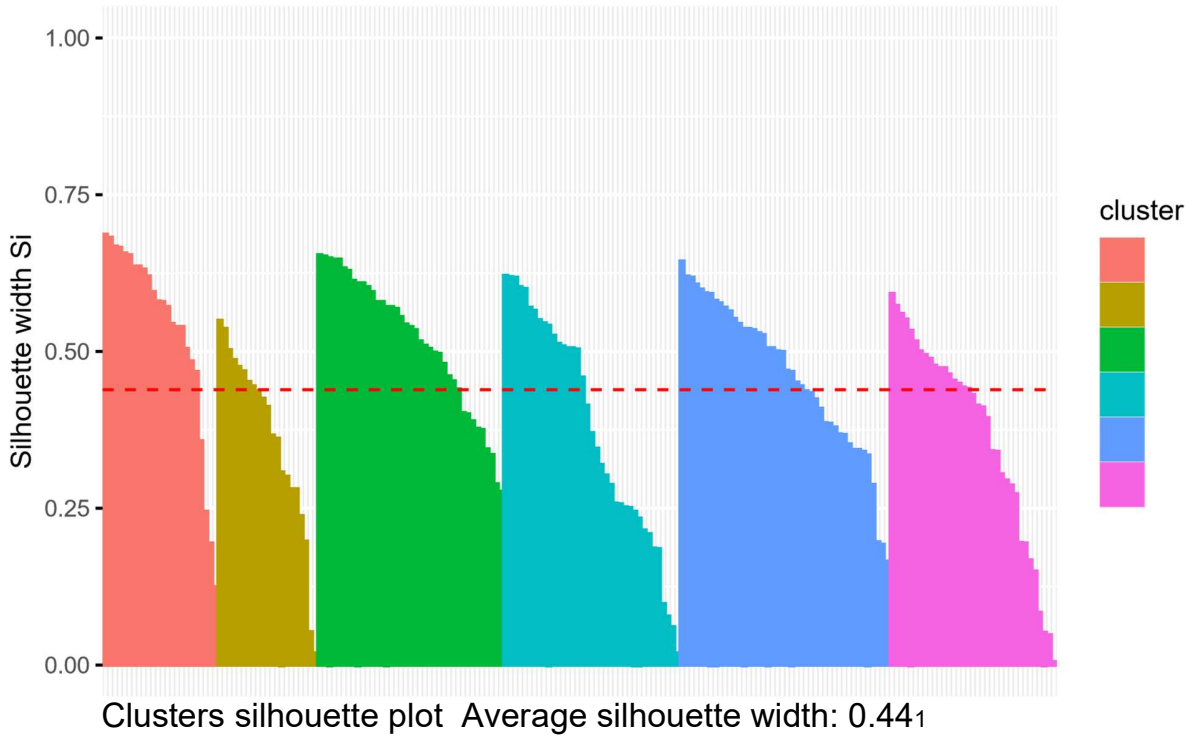


Number of Clusters

```
#Grupujemy nasze dane z podziałem na 6 grup i dokonujemy wizualizacji na wykresie
 set.seed(1234)
km <- kmeans(dane_std,6, iter.max = 100)
 grupy_km<-data.frame(dane,km$cluster)
fviz_cluster(km, data = dane_std)
```

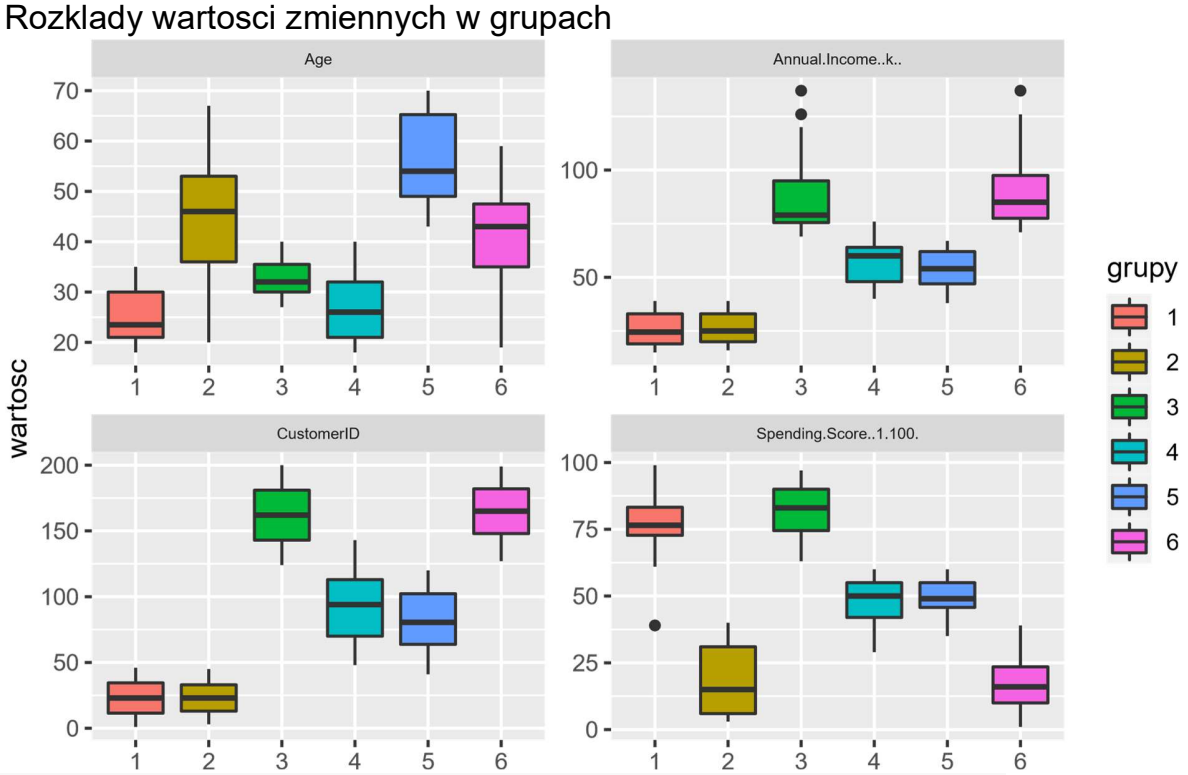



```
#Żeby upewnić się, że 6 grup jest dobrym wyborem tworzę wykres sylwetkowy
#Jak widzimy na wykresie sylwetkowym nie ma żadnych wartośći poniżej 0 i średnia
#wartość sylwetki jest większa niż w przypadku gdy wybralibyśmy 4 grupy więc 6
#jest lepszą liczbą grup
d <- dist(dane_std, method = "euclidean")
fviz_silhouette(silhouette(km$cluster, d))
```




```
#Dokonam teraz opisu wartości zmiennych w poszczególnych grupach dane1 %>%
   mutate(cluster=as.factor(km$cluster))%>% group_by(cluster)%>%

pivot_longer(1:4)%>% ggplot(aes(x=cluster,y=value, fill = cluster))+geom_boxplot()+
facet_wrap(vars(name),scales="free")+ labs(x="",y="wartość",fill="grupy",title="Rozkłady wartości
zmiennych w grupach")+ theme(strip.text = element_text(size=6))
```

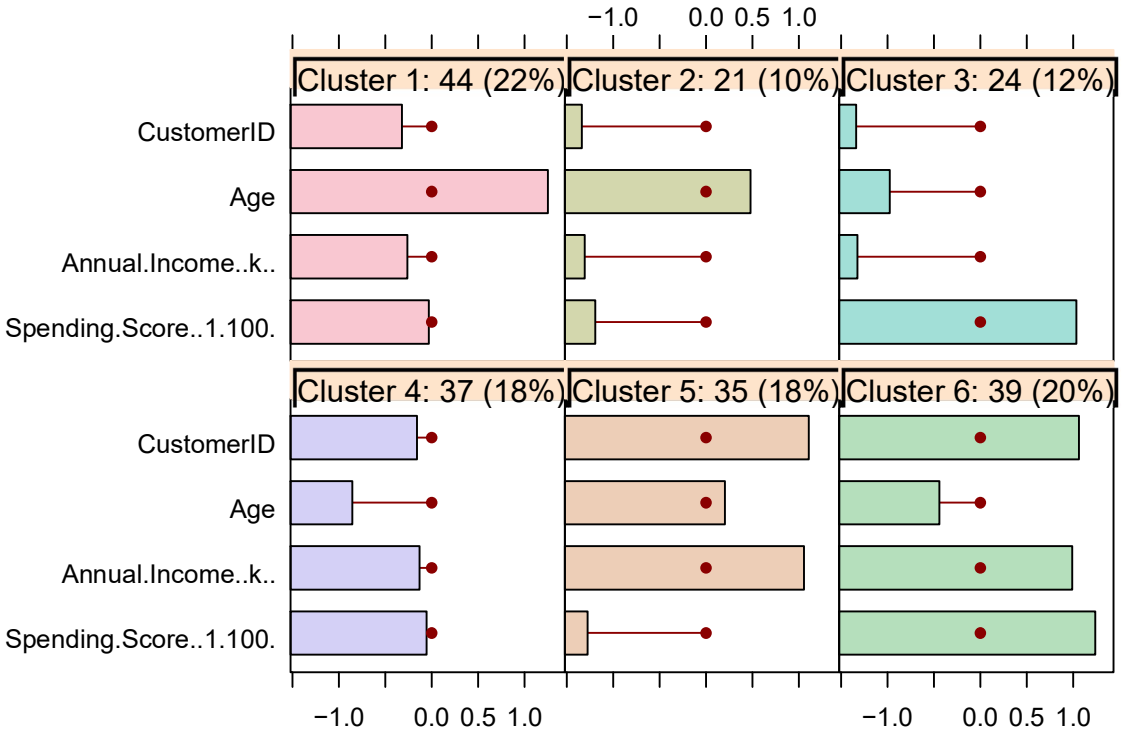


#grupa 1 charakteryzuje się niskim wiekiem (od 18 do 35 lat) i niskim przychodem
#ale wysoką punktacją wydatków
#grupa 2 to najbardziej zróżnicowana pod względem wieku grupa osoby w niej
#posiadają niski poziom przychodów oraz niską punktację wydatków
#grupa 3 to osoby około 30 roku życia z wysokimi przychodami i wysoką punktacją
#wydatkóW
#grupa 4 to osoby po 20 roku życia do około 40 roku życia ich przychód jest na #średnim poziomie tak samo jak ich puntkacja wydatków #grupa 5 to najstarsza ze wszystkich grup jej przychody są na podobnym poziomie
 # co grupy 4 tak samo jak punktacja wydatków
#grupa 6 to bardzo zróżnicowana pod wzgledem wieku grupa mogą się w niej znaleźć #zarówno osoby mające lat 20 jak i lat 55 charakteryzuje się wysokim poziomem #przychodów ale niską punktacją wydatków.

```
MD.km28 <- relabel(MD.km28) MD.k4 <- MD.km28[["6"]]
MD.k4


barchart(MD.k4, shade = TRUE)


#Na poniższym wykresie widzimy zależności opisane wyżej ale przedstawione w inny sposób
```


−1.0 0.0 0.5 1.0
Cluster 1: 44 (22%)
Cluster 2: 21 (10%)
Cluster 3: 24 (12%)
CustomerID
Age
Annual.Income..k..
Spending.Score..1.100.
Cluster 4: 37 (18%)
Cluster 5: 35 (18%)
Cluster 6: 39 (20%)
CustomerID
Age
Annual.Income..k..
Spending.Score..1.100.
−1.0 0.0 0.5 1.0
−1.0 0.0 0.5 1.0