```
# Подключаем библиотеку ISLR, где находится таблица Credit

library(ISLR)

x1<-data.frame(cbind(Income=Credit$Income, Rating=Credit$Rating, Student=Credit$Student, Balance=Credit$Balance))

set.seed(1)

# Формируем четыре кластера

myclasters <- kmeans(x = x1, centers = 4)

myclasters$cluster

# Центры кластеров

myclasters$centers



cluster1<-Credit[myclasters$cluster==1,-1]

cluster2<-Credit[myclasters$cluster==2,-1]

cluster3<-Credit[myclasters$cluster==3,-1]

cluster4<-Credit[myclasters$cluster==4,-1]

#Для проверки мультиколлинеарности с помощью функции vif

#подключаем библиотеку car

library(car)


#Исследуем первый кластер

#Исследуем модель линейной множественной регрессии,

#без учета качественных переменных

model_1<-lm(data=cluster1[,-c(7:10)], Balance~.)

summary(model_1)

vif(model_1)

#Исключаем переменную Limit, как сильно зависимую от остальных признаков

model_1<-lm(data=cluster1[,-c(7:10)], Balance~.-Limit)

vif(model_1)

model_1<-lm(data=cluster1, Balance~.-Limit-Income)
```

```
vif(model_1)
summary(model_1)
# включаем в итоговую модель только значимые переменные
model_1<-lm(data=cluster1, Balance~Rating+Age+Student)
summary(model_1)
library(ggplot2)
ggplot(data=cluster1, aes(x=Rating, y=Balance, colour = Student))+geom_point()+
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE,
              aes(color=Student))

#Исследуем второй кластер
model_2<-lm(data=cluster2[,-c(7:10)], Balance~.)
summary(model_2)
vif(model_2)
model_2<-lm(data=cluster2[,-c(7:10)], Balance~.-Limit)
vif(model_2)
model_2<-lm(data=cluster2, Balance~.-Limit)
summary(model_2)
# Включаем в модель только значимые переменные
model_2<-lm(data=cluster2, Balance~Income+Rating+Age+Student+Married)
summary(model_2)

#Исследуем третий кластер
model_3<-lm(data=cluster3[,-c(7:10)], Balance~.)
summary(model_3)
vif(model_3)
model_3<-lm(data=cluster3[,-c(7:10)], Balance~.-Limit)
vif(model_3)
model_3<-lm(data=cluster3, Balance~.-Limit)
summary(model_3)
```

```
# Включаем в модель только значимые переменные
model_3<-lm(data=cluster3, Balance~Income+Rating+Age+Student)
summary(model_3)

boxplot(cluster3$Balance~cluster3$Student)

#Исследуем четвертый кластер с самой маленькой задолженностью
model_4<-lm(data=cluster4[,-c(7:10)], Balance~.)
vif(model_4)
model_4<-lm(data=cluster4[,-c(7:10)], Balance~.-Limit)
vif(model_4)
model_4<-lm(data=cluster4, Balance~.-Limit)
summary(model_4)
model_4<-lm(data=cluster4, Balance~Income+Rating+Student)
summary(model_4)

boxplot(cluster4$Balance~cluster4$Student)
# Много нулевых значений баланса
ggplot(data=cluster4, aes(x=Rating, y=Balance, colour = Student))+geom_point()+
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE,
              aes(color=Student))

#Более детальное исследование кластера 4
# Много нулевых значений у переменной Balance
# Итоговая модель для наблюдений, у которых рейтинг превышает 210
model_4_1<-lm(data=cluster4[cluster4$Rating>210,], Balance~Income+Rating+Student)
summary(model_4_1)
# Итоговая модель для студентов, у которых рейтинг меньше 210
model_4_2<-lm(data=cluster4[cluster4$Rating<=210 & cluster4$Student=="Yes",],
```

```
        Balance~Income+Rating)
summary(model_4_2)
#Баланс равен нулю для не студентов с рейтингом, меньшим 210
cluster4[cluster4$Rating<=210 & cluster4$Student=="No",]$Balance

#Средние значения переменной Age
mean(cluster4$Age)
mean(cluster1$Age)
mean(cluster2$Age)
mean(cluster3$Age)
```