Submission 2

Nicholas Sherman, Kishor Kumar, Sridhar, Adjoa Adanledji, and Andrew Smith

11/18/2019

Data Cleaning

```
cData$CustomerId = as.character(cData$CustomerId)
cData$CreditScore = as.integer(cData$CreditScore)
cData$CreditScoreBins = cut(x = cData$CreditScore, breaks = c(349, seq(360,850,10)))
cData$Geography = as.factor(cData$Geography)
cData$Gender = as.factor(cData$Gender)
cData$Age = as.integer(cData$Age)
cData$AgeBins = cut(x = cData$Age, breaks = c(17, seq(25,100,5)))
cData$TenureFactor = factor(as.character(cData$Tenure), levels = c("0","1","2","3","4","5","6","7","8",
cData$Tenure = as.factor(as.character(cData$Tenure))
cData$NumOfProducts = as.integer(cData$NumOfProducts)
cData$HasCrCard = as.factor(as.character(cData$HasCrCard))
levels(cData$HasCrCard) = c("inactive","active")
cData$IsActiveMember = as.factor(as.character(cData$IsActiveMember)) #assuming 1 = member, 0 = not
levels(cData$IsActiveMember) = c("inactive","active")
cData$Exited = as.factor(as.character(cData$Exited))
levels(cData$Exited) = c("retained","exited") #0 = retained, 1 = exited
cData = cData %>% select(-RowNumber)
```

Data Exploration

Difference between Active and Inactive

```
##           IsActiveMember
## Exited     inactive active
##   retained     3547   4416
##   exited       1302    735
```

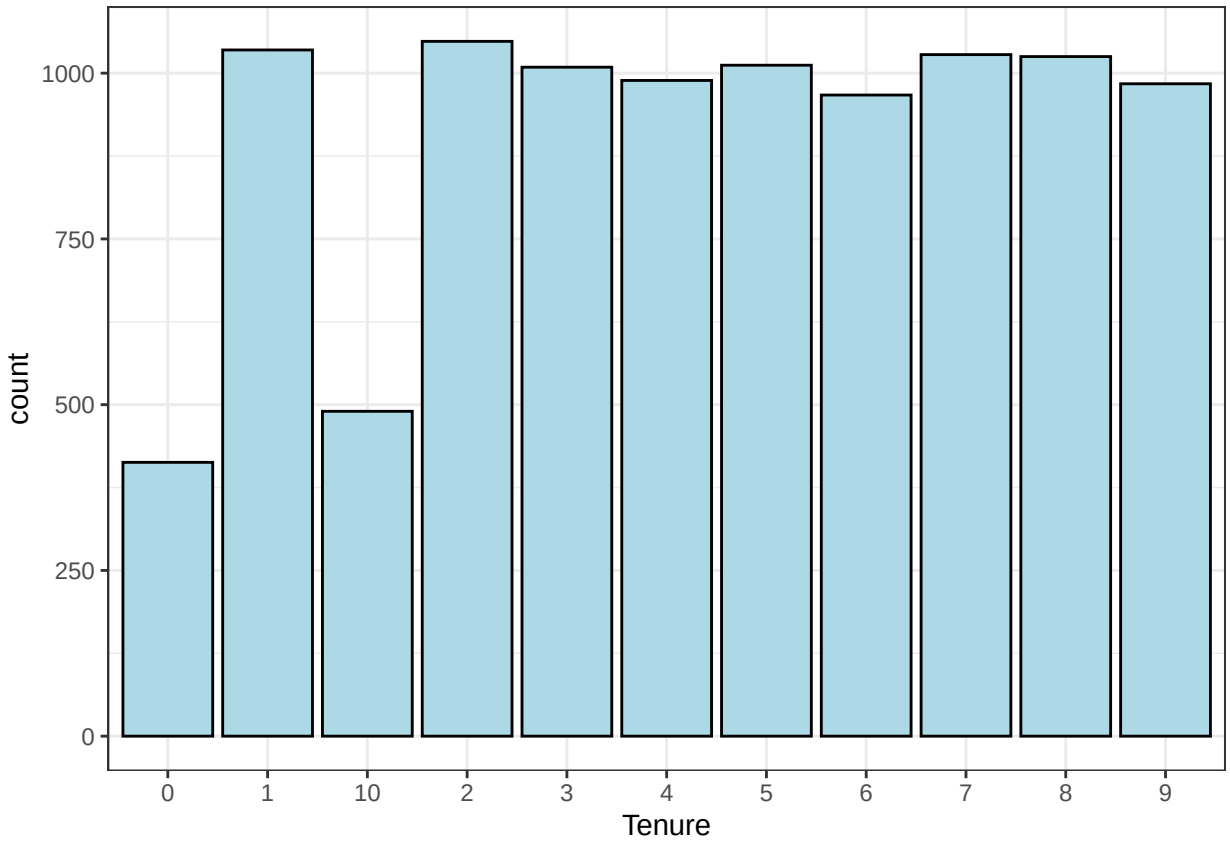
```
## [1] 10000
```

```
## [1] 2932
```

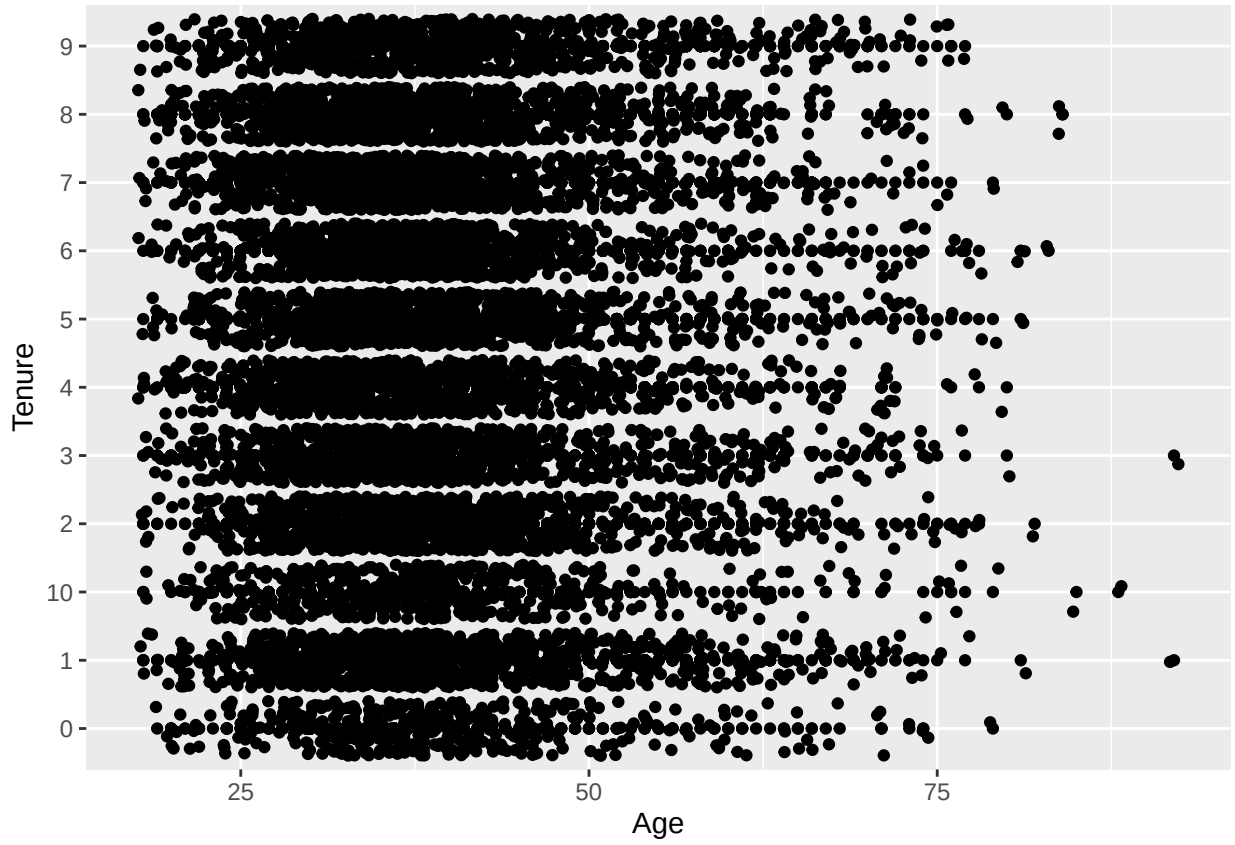
```
## [1] 7068
```

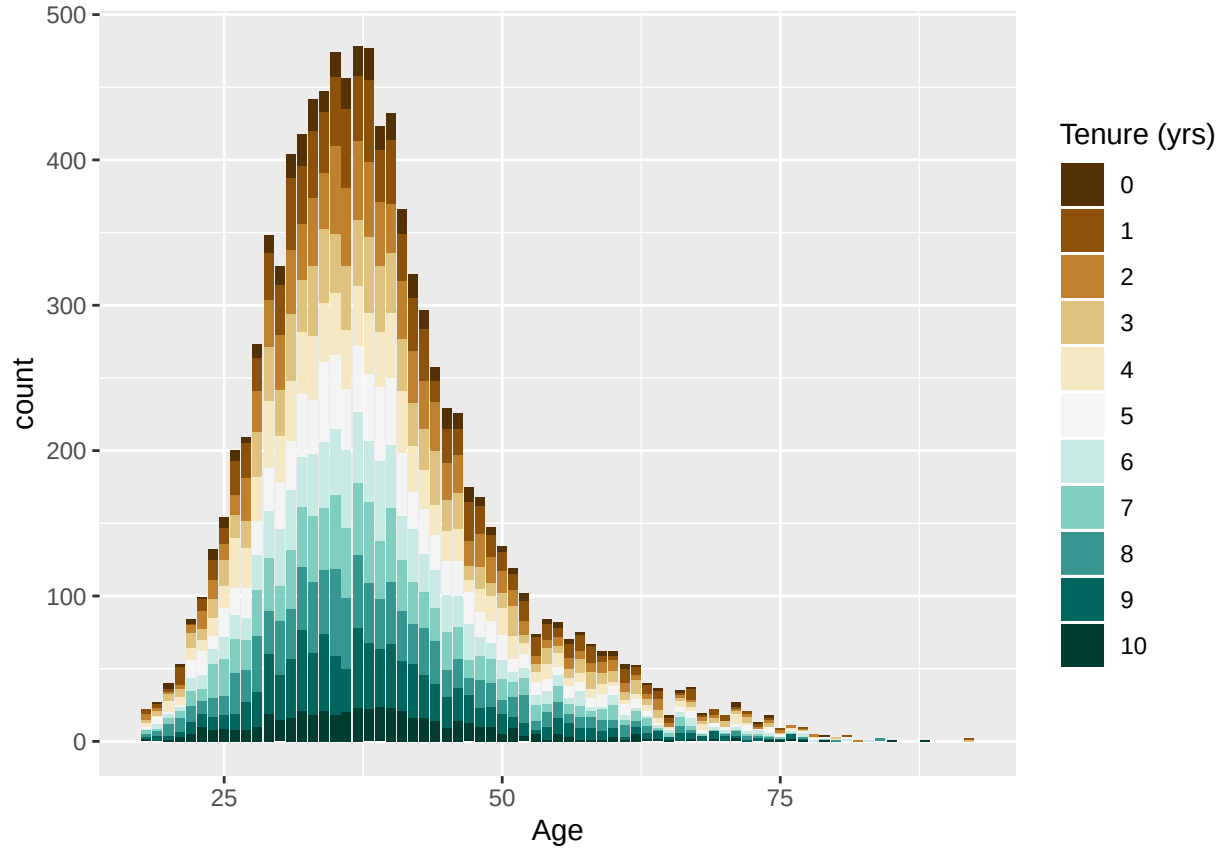
Possiblity of duplicate values.

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```


1000
750
500
250
0
count
0
1
10
2
3
4
5
6
7
8
9
Tenure

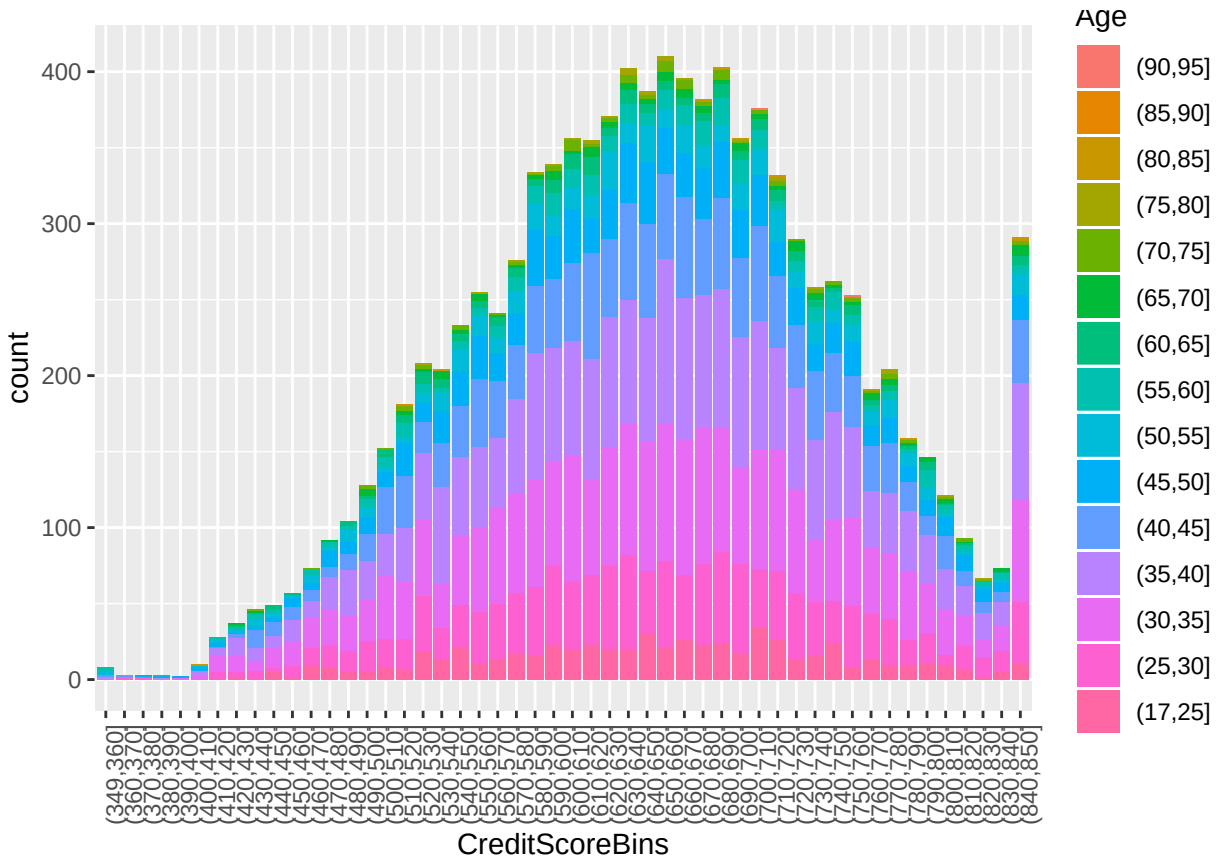
How do age and tenure compare?



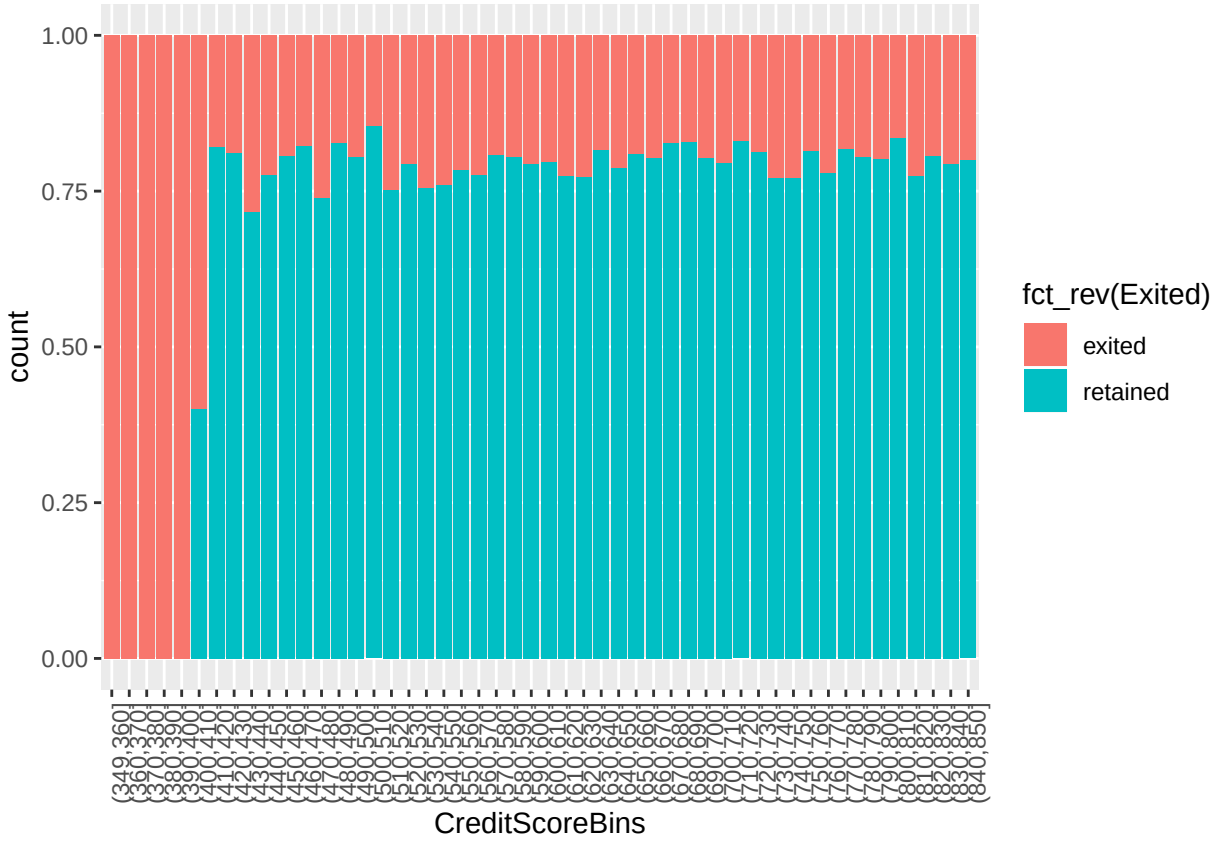


Looking into age groups

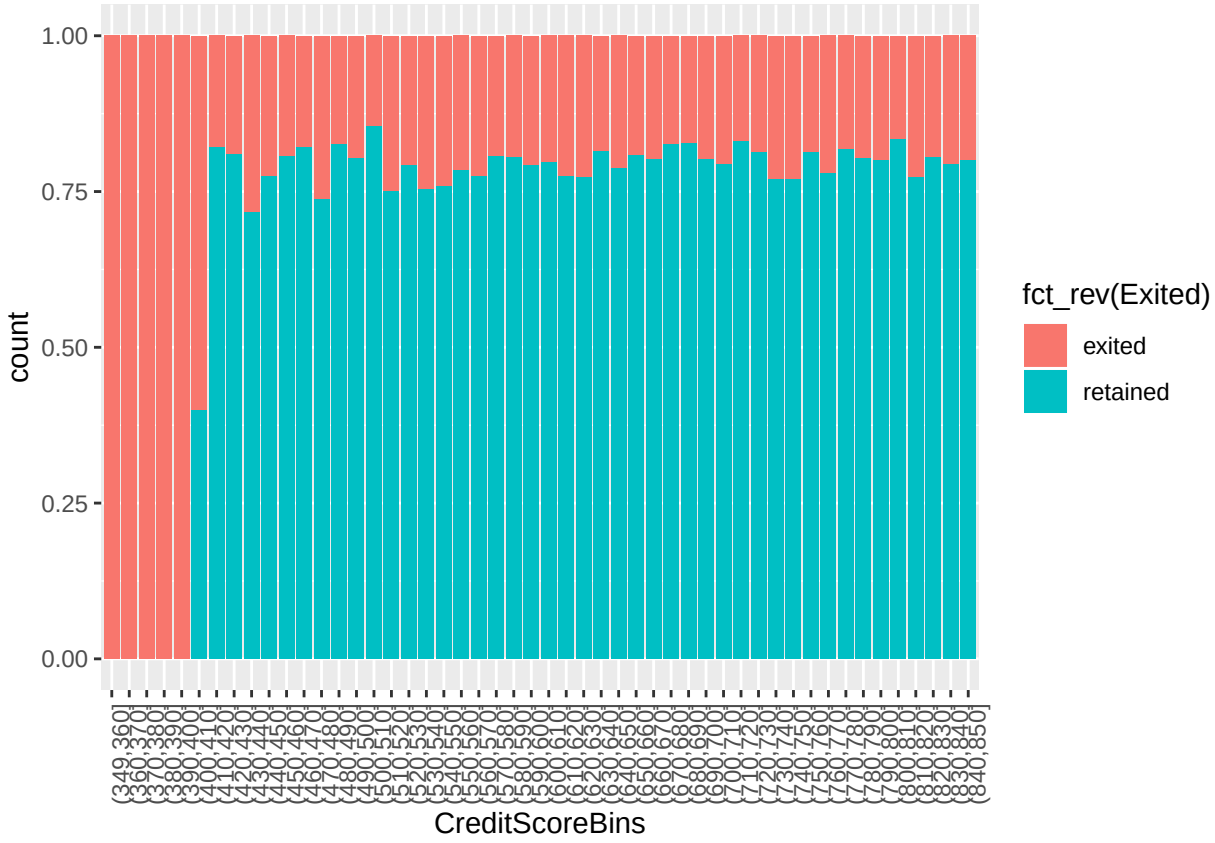
```
##
##  (17,25]  (25,30]  (30,35]  (35,40]  (40,45]  (45,50]  (50,55]  (55,60]
##      611     1357     2185     2266     1470      850      461      336
##  (60,65]  (65,70]  (70,75]  (75,80]  (80,85]  (85,90]  (90,95] (95,100]
##      200      131       88       33        9        1        2        0
```



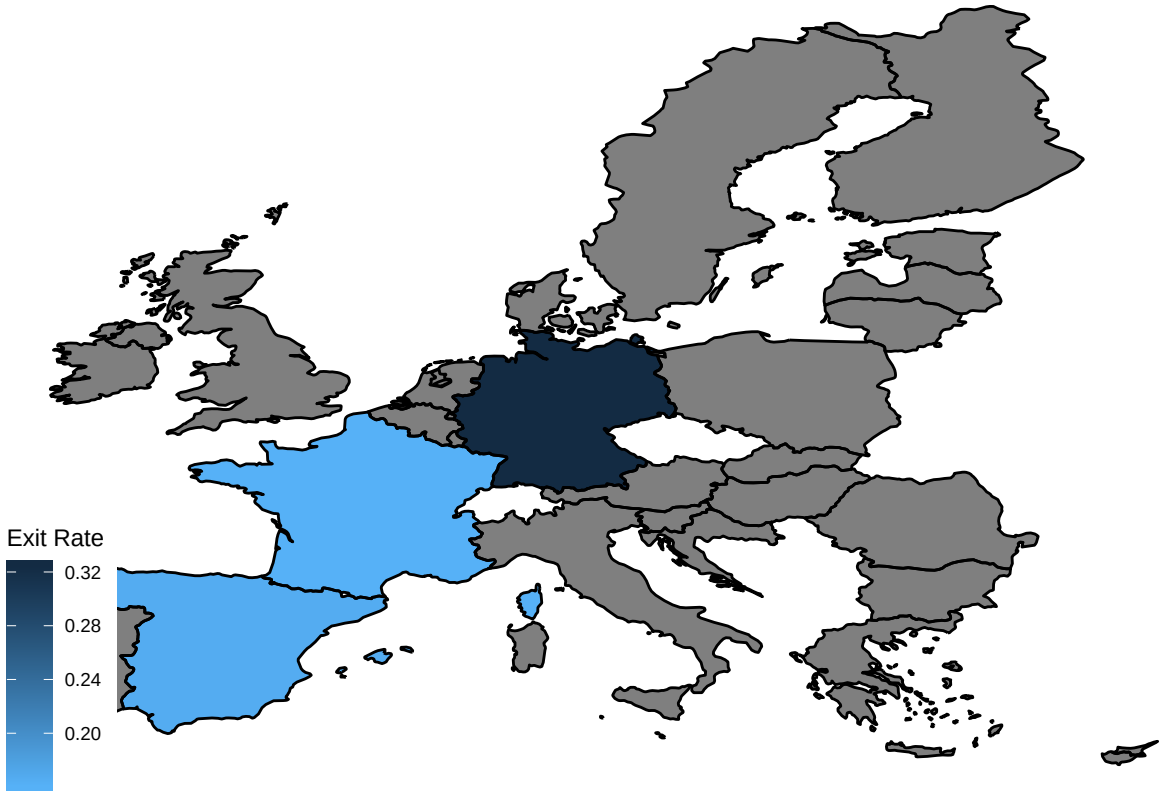
Data Exploration Response Variable: Exited



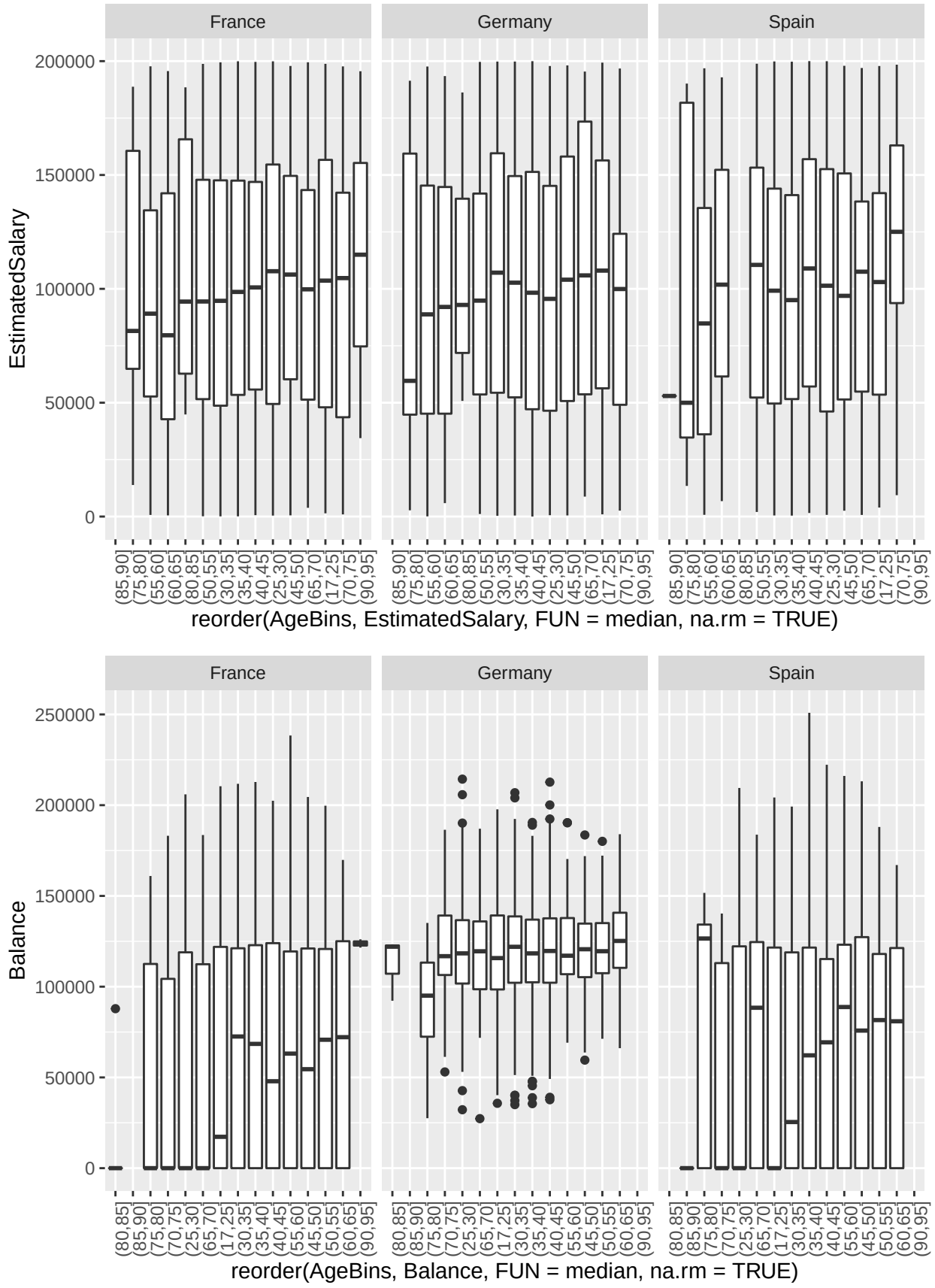
A low credit score maybe implies the bank didn't even accept them, or high fees and the customer didn't want to stay.


1.00
0.75
0.50
0.25
0.00
count
(349,360]
(360,370]
(370,380]
(380,390]
(390,400]
(400,410]
(410,420]
(420,430]
(430,440]
(440,450]
(450,460]
(460,470]
(470,480]
(480,490]
(490,500]
(500,510]
(510,520]
(520,530]
(530,540]
(540,550]
(550,560]
(560,570]
(570,580]
(580,590]
(590,600]
(600,610]
(610,620]
(620,630]
(630,640]
(640,650]
(650,660]
(660,670]
(670,680]
(680,690]
(690,700]
(700,710]
(710,720]
(720,730]
(730,740]
(740,750]
(750,760]
(760,770]
(770,780]
(780,790]
(790,800]
(800,810]
(810,820]
(820,830]
(830,840]
(840,850]
CreditScoreBins
fct_rev(Exited)
exited
retained

Geography

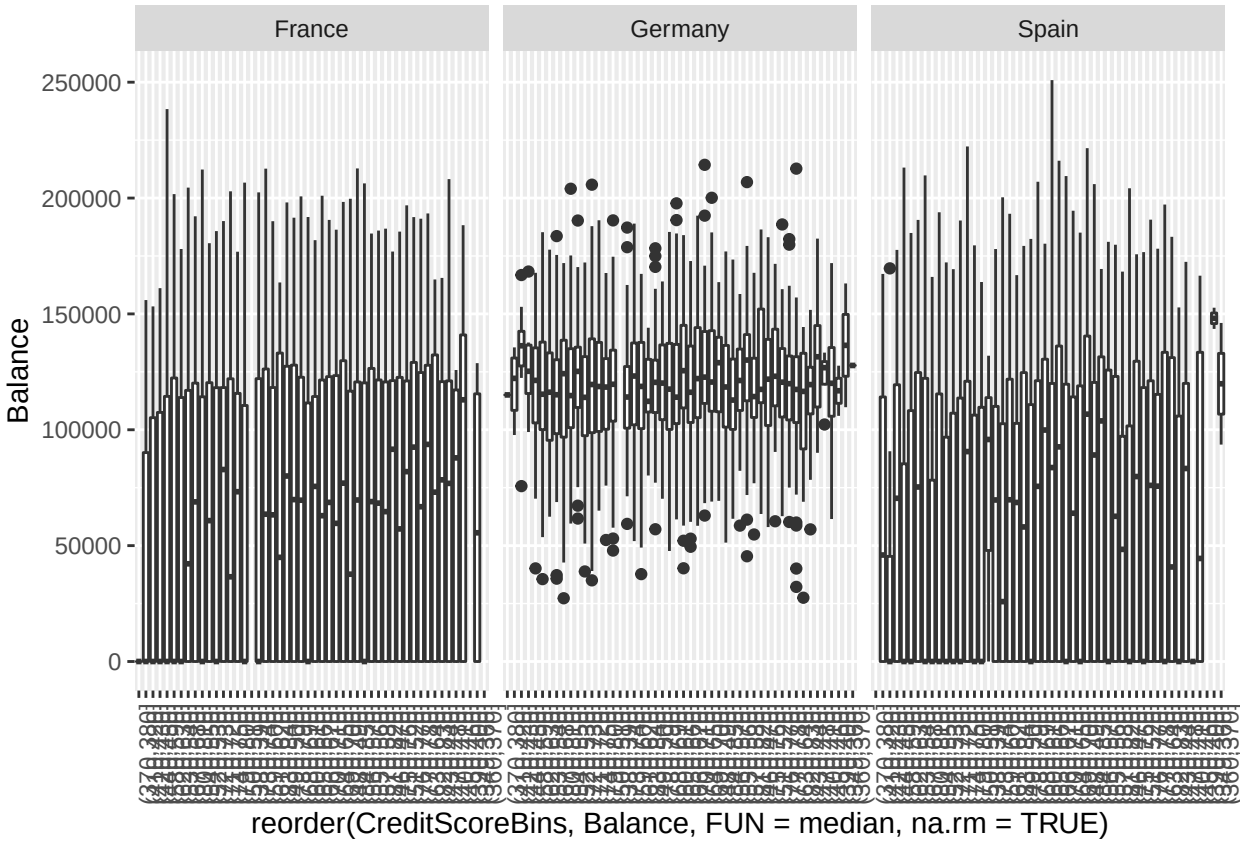


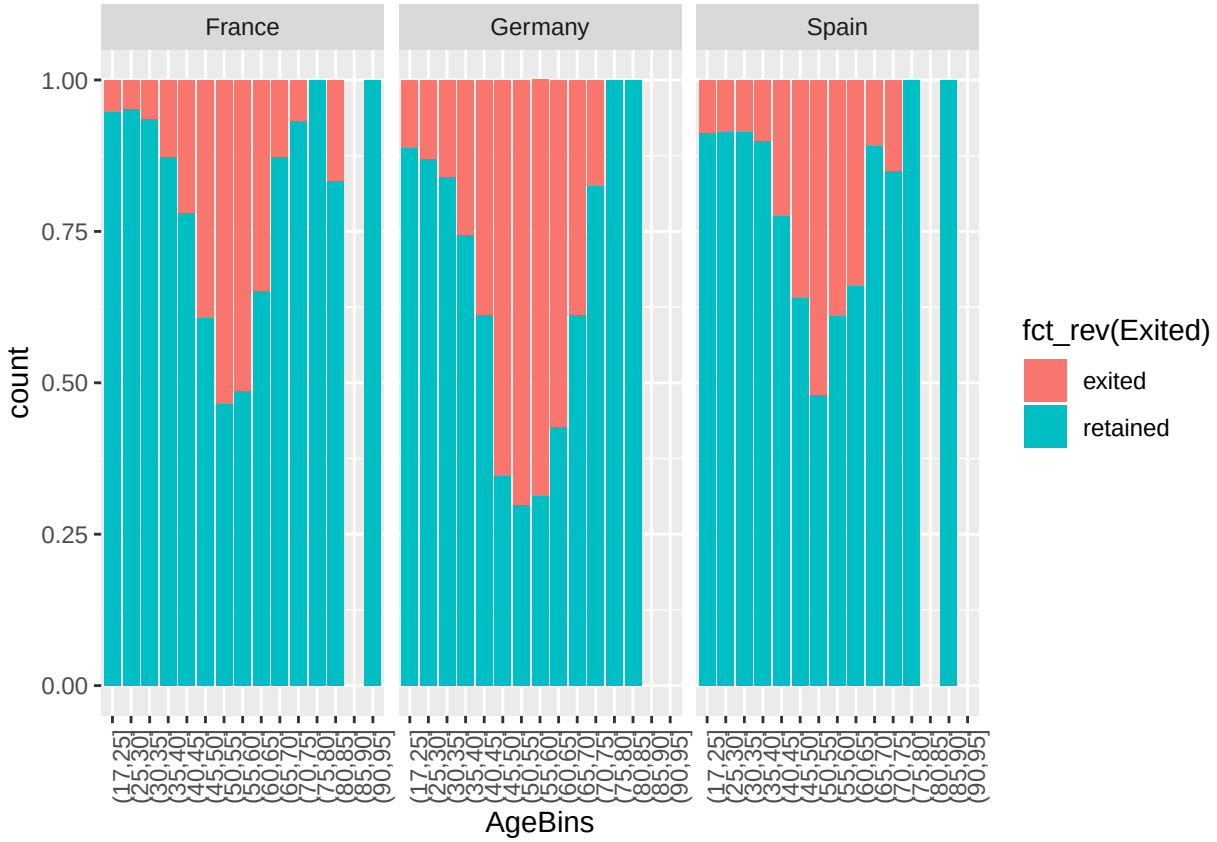
Germany

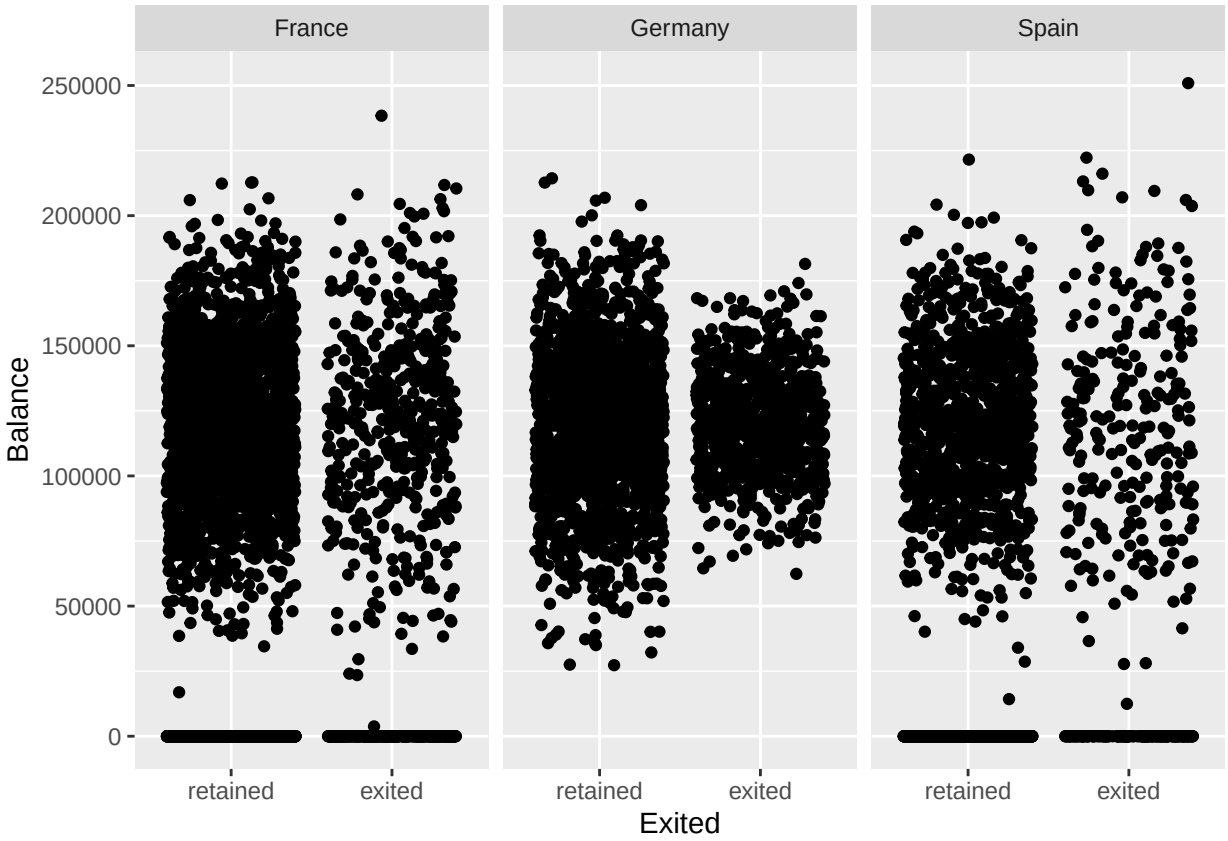

France
Germany
Spain
EstimatedSalary
200000
150000
100000
50000
0
(85,90]
(75,80]
(55,60]
(60,65]
(80,85]
(50,55]
(30,35]
(35,40]
(40,45]
(25,30]
(45,50]
(65,70]
(17,25]
(70,75]
(90,95]
reorder(AgeBins, EstimatedSalary, FUN = median, na.rm = TRUE)
France
Germany
Spain
Balance
250000
200000
150000
100000
50000
0
(80,85]
(85,90]
(75,80]
(70,75]
(25,30]
(65,70]
(17,25]
(30,35]
(35,40]
(40,45]
(55,60]
(45,50]
(50,55]
(60,65]
(90,95]
reorder(AgeBins, Balance, FUN = median, na.rm = TRUE)

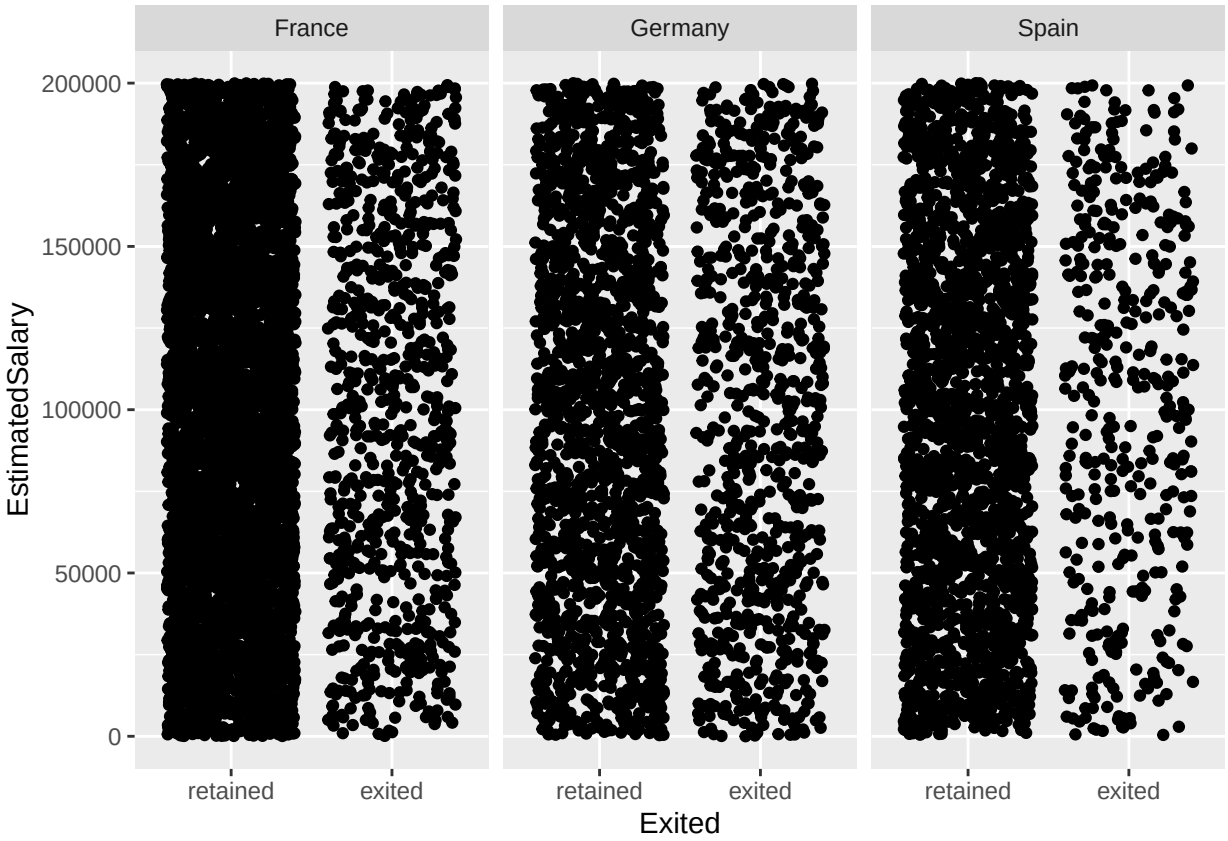
A lot of outliers are present in Germany compared to the other two countries. Same is the case if we create a boxplot with respect to balance and credit score bins.

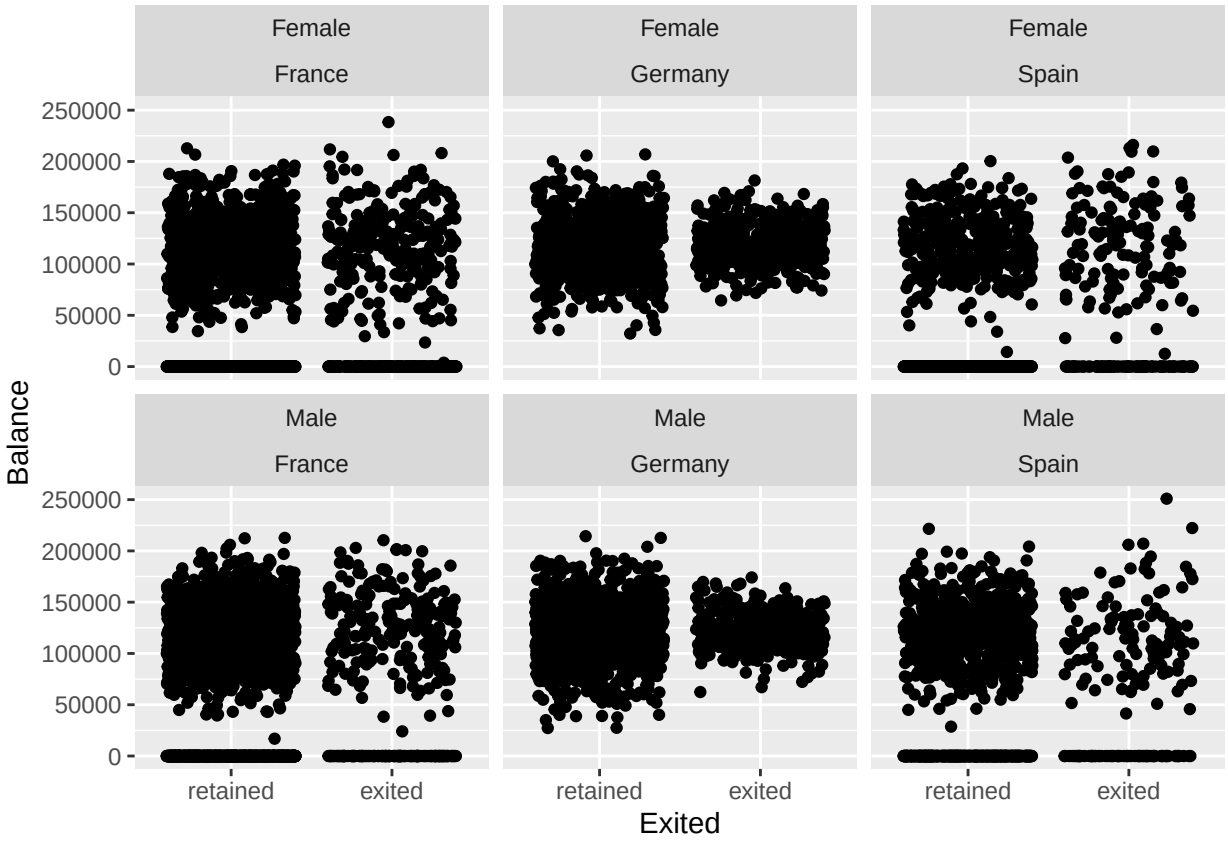
More Geography

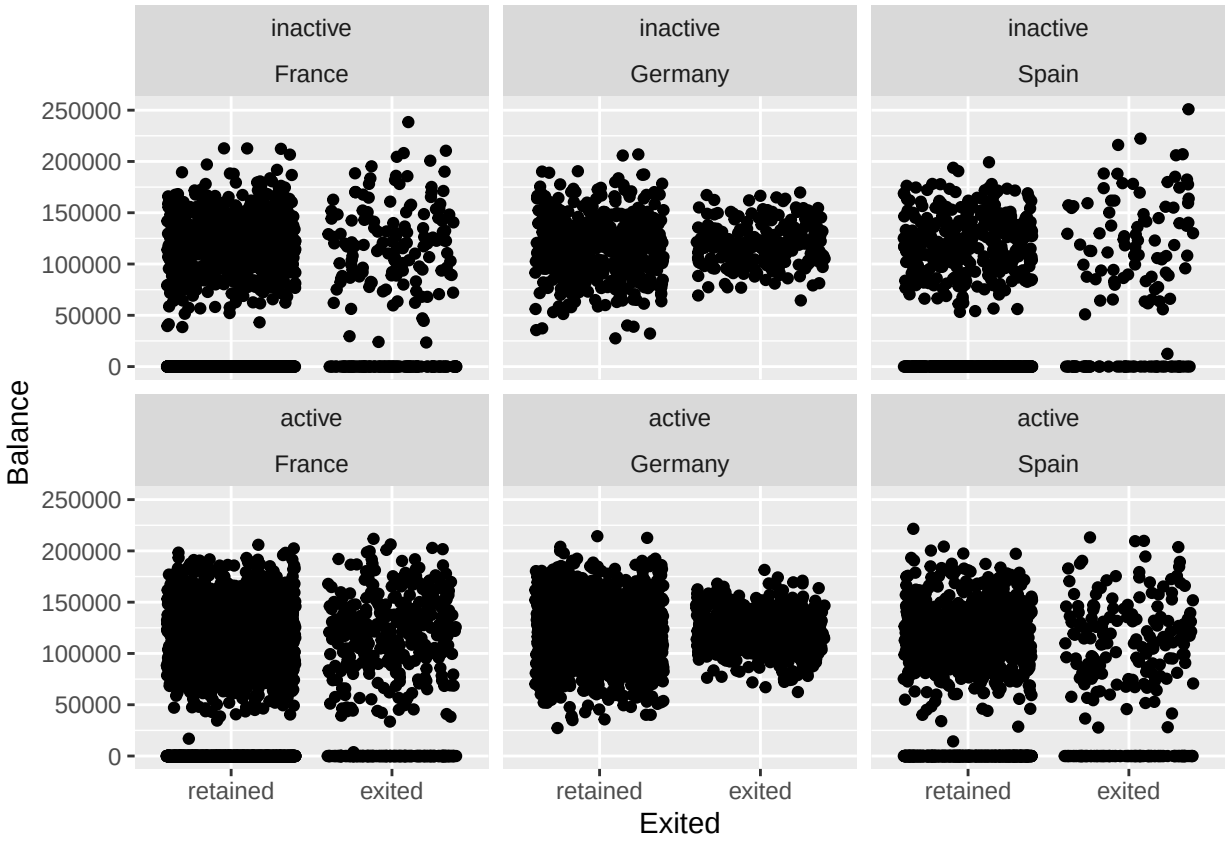




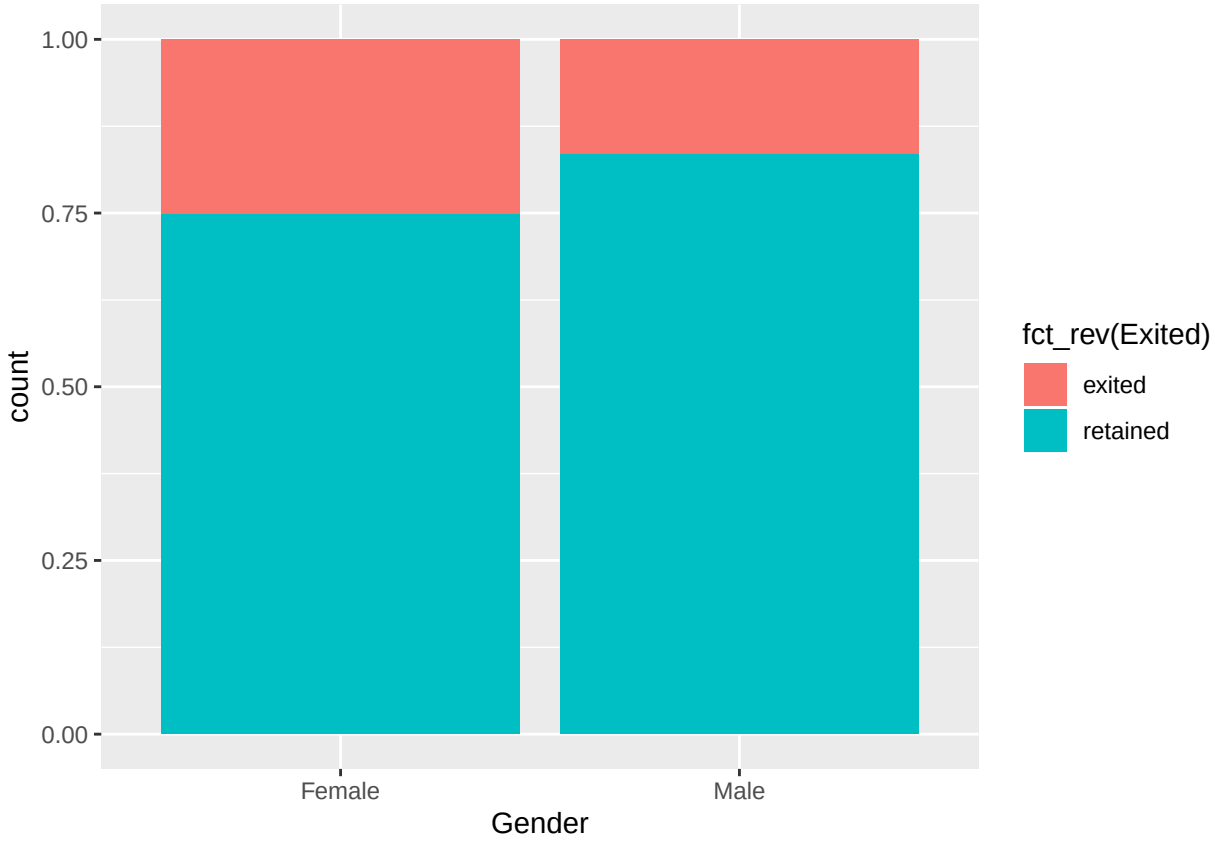

France
Germany
Spain
250000
200000
150000
100000
50000
0
Balance
retained
exited
retained
exited
retained
exited
Exited


France
Germany
Spain
200000
150000
100000
50000
0
EstimatedSalary
retained
exited
retained
exited
retained
exited
Exited

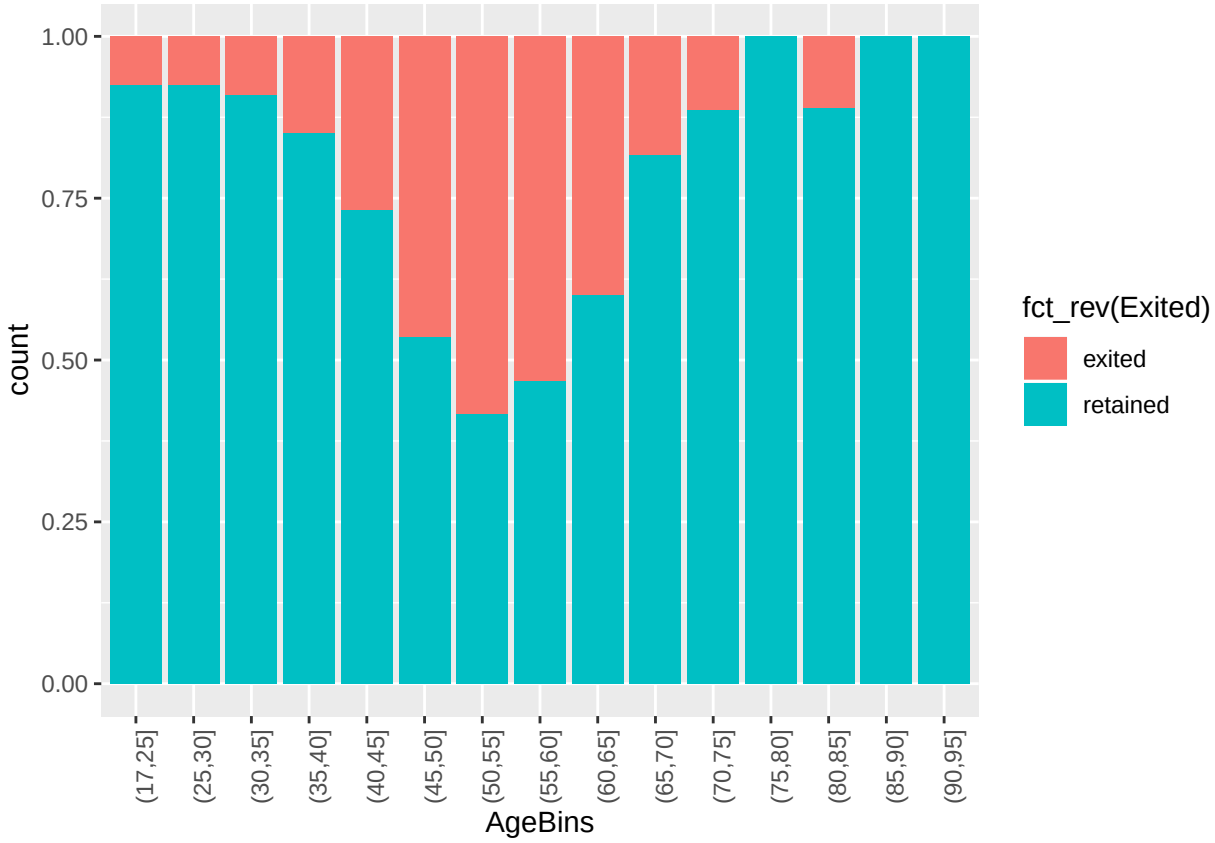

Female
France
Female
Germany
Female
Spain
Male
France
Male
Germany
Male
Spain
Balance
250000
200000
150000
100000
50000
0
retained
exited
Exited



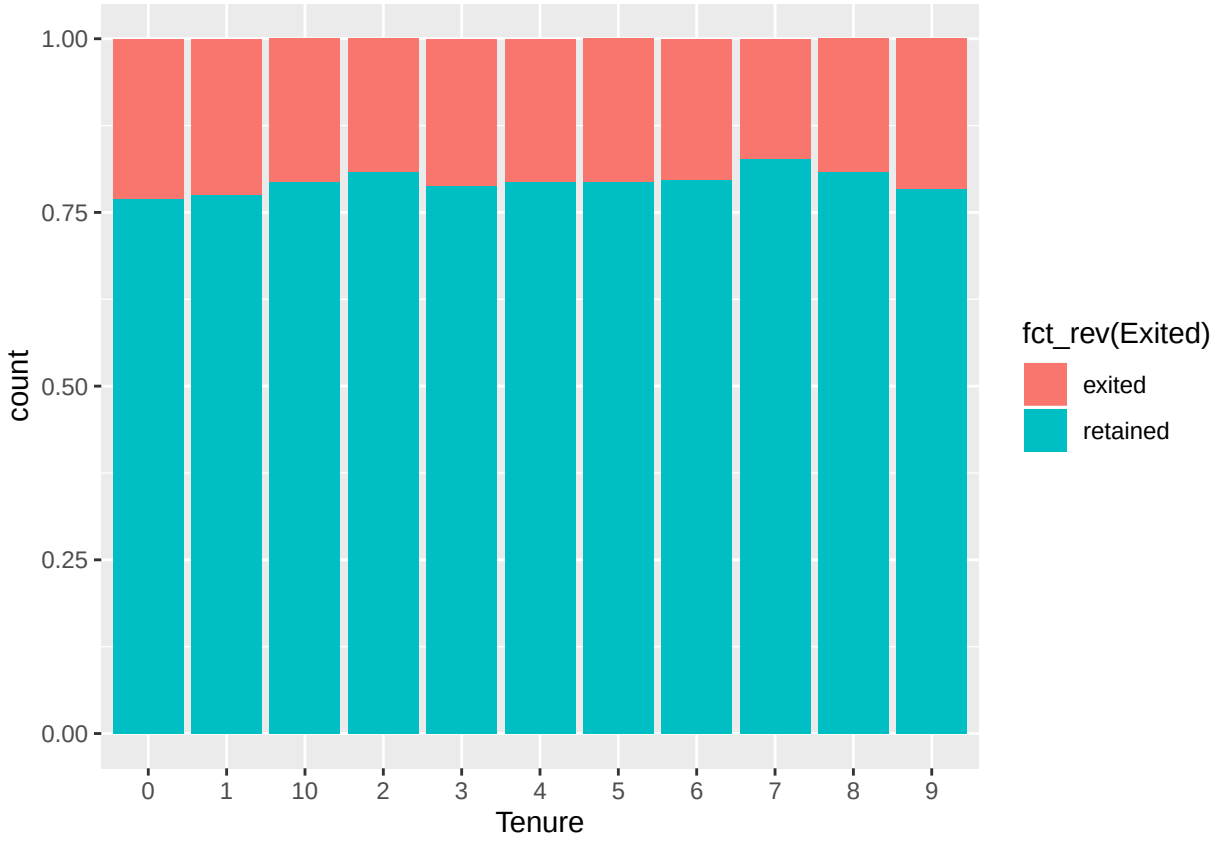
There doesn't seem to be much explanation with respect to the people who exited and their estimated salary However, the spread of "balance" of the people who exited in Germany seems to be less and is hovering around 75000 to 175000, and this seems to be true for both the genders. But the number of exited people in Germany seems to be considerably higher for customers with Credit Card than without it comparing to the numbers in the other two countries ### GENDER


1.00
0.75
0.50
0.25
0.00
count
Female
Male
Gender
fct_rev(Exited)
exited
retained

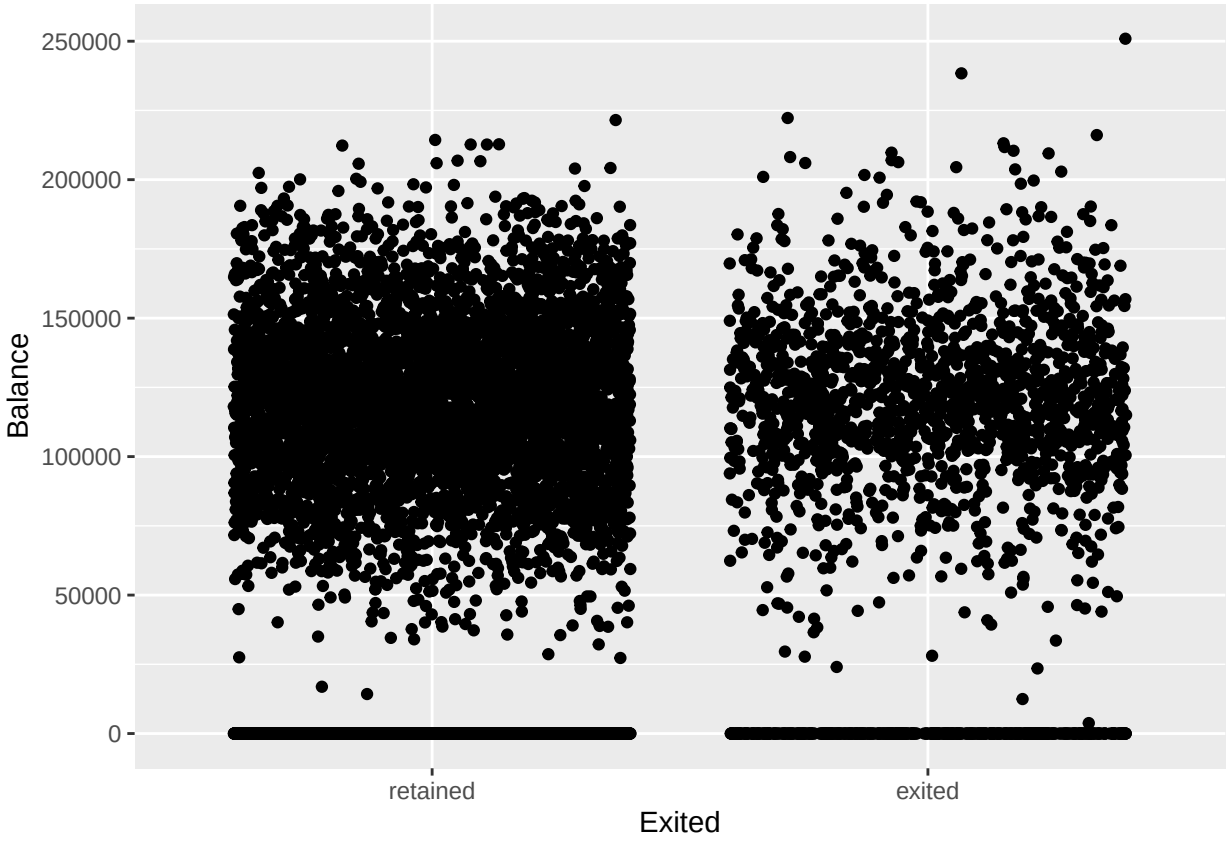
AGE

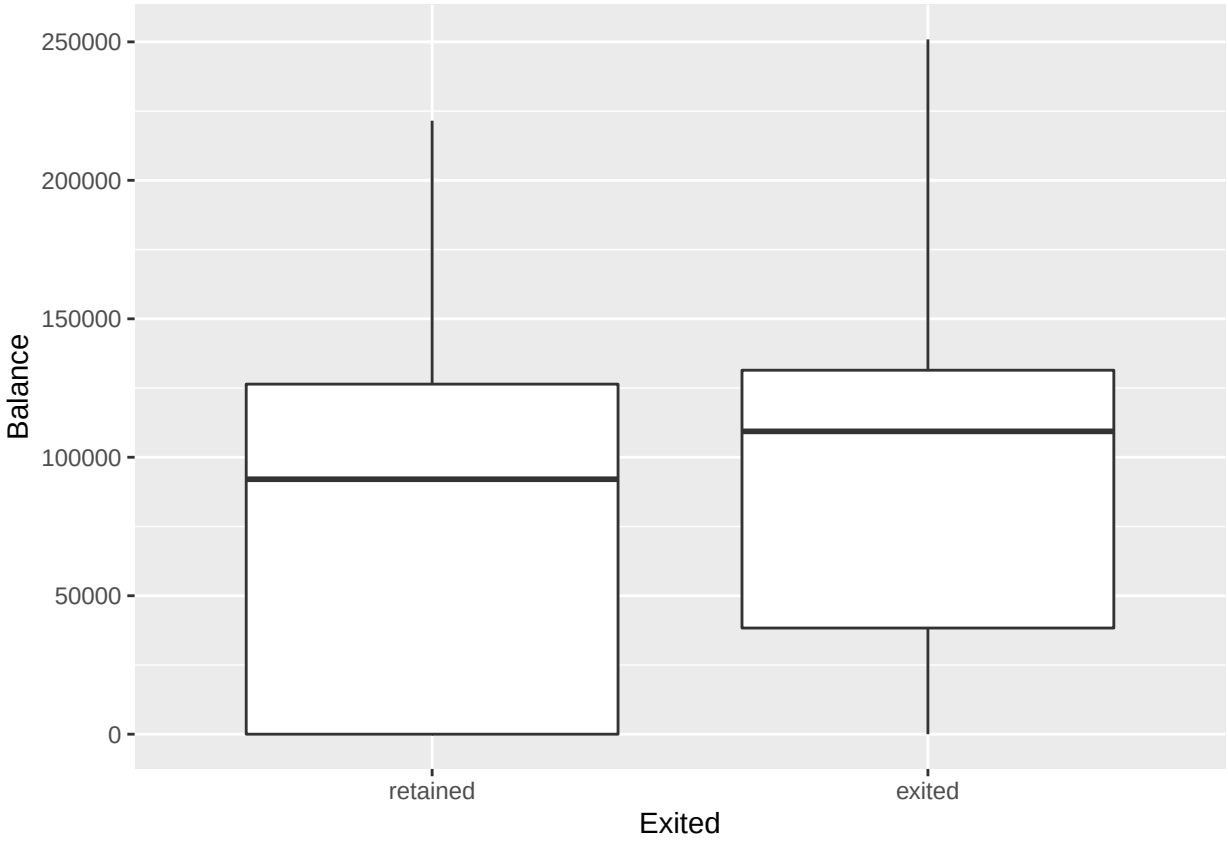

1.00
0.75
0.50
0.25
0.00
count
(17,25]
(25,30]
(30,35]
(35,40]
(40,45]
(45,50]
(50,55]
(55,60]
(60,65]
(65,70]
(70,75]
(75,80]
(80,85]
(85,90]
(90,95]
AgeBins
fct_rev(Exited)
exited
retained

TENURE

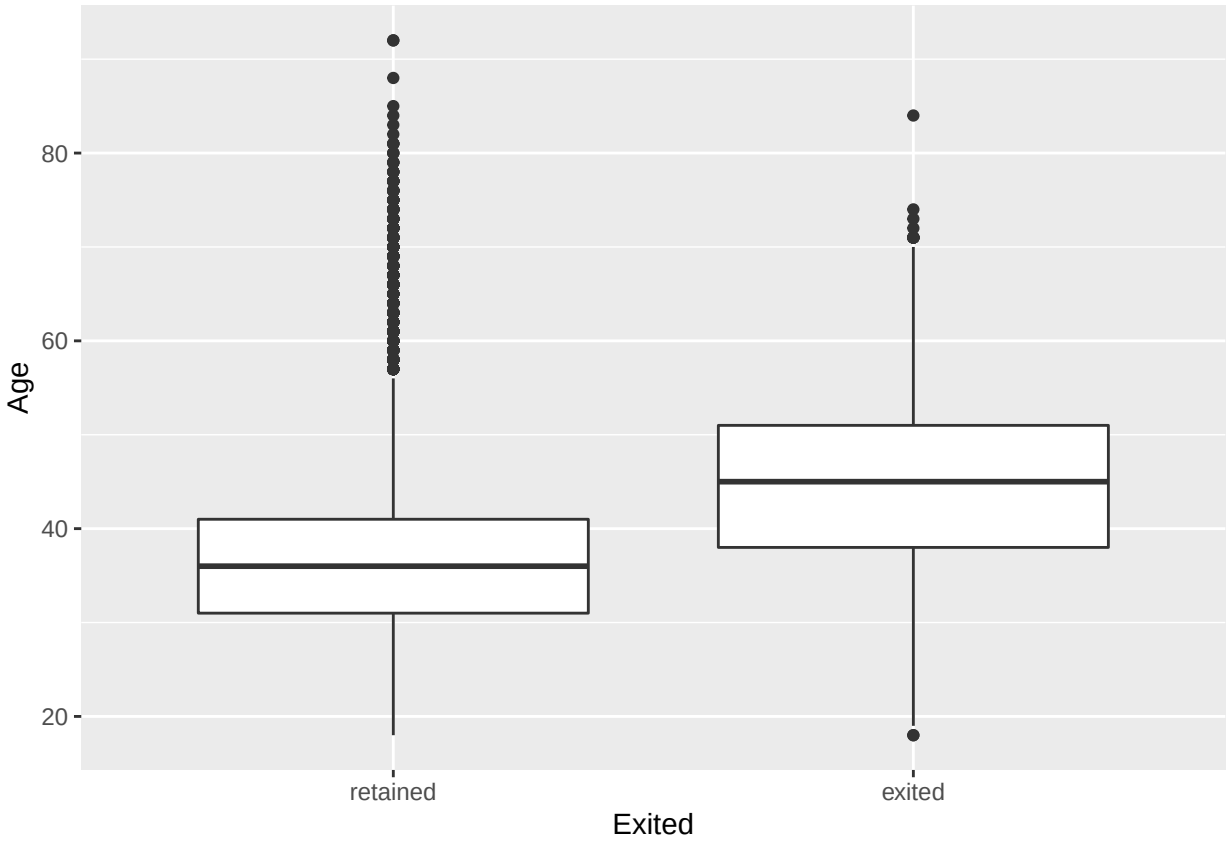

1.00
0.75
0.50
0.25
0.00
count
0
1
10
2
3
4
5
6
7
8
9
Tenure
fct_rev(Exited)
exited
retained

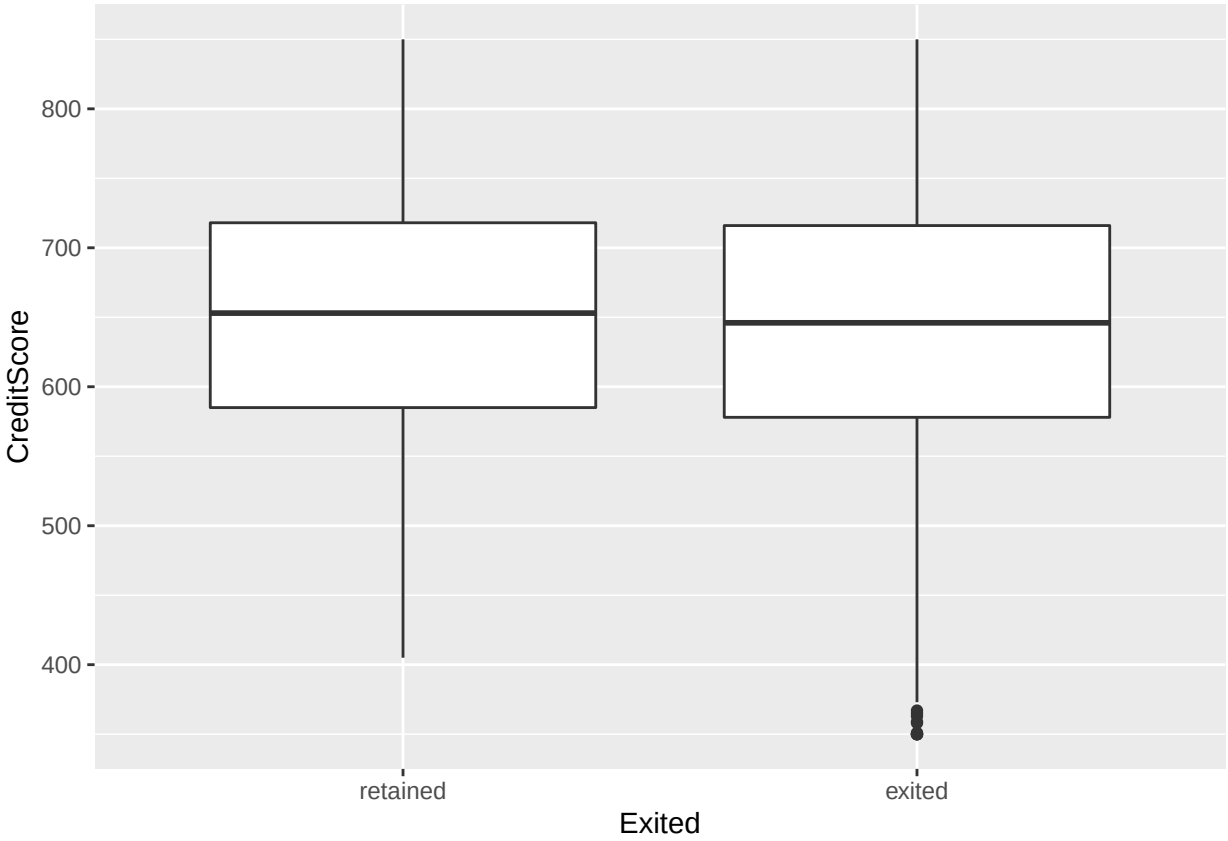
BALANCE


250000
200000
150000
100000
50000
0
Balance
retained
exited
Exited

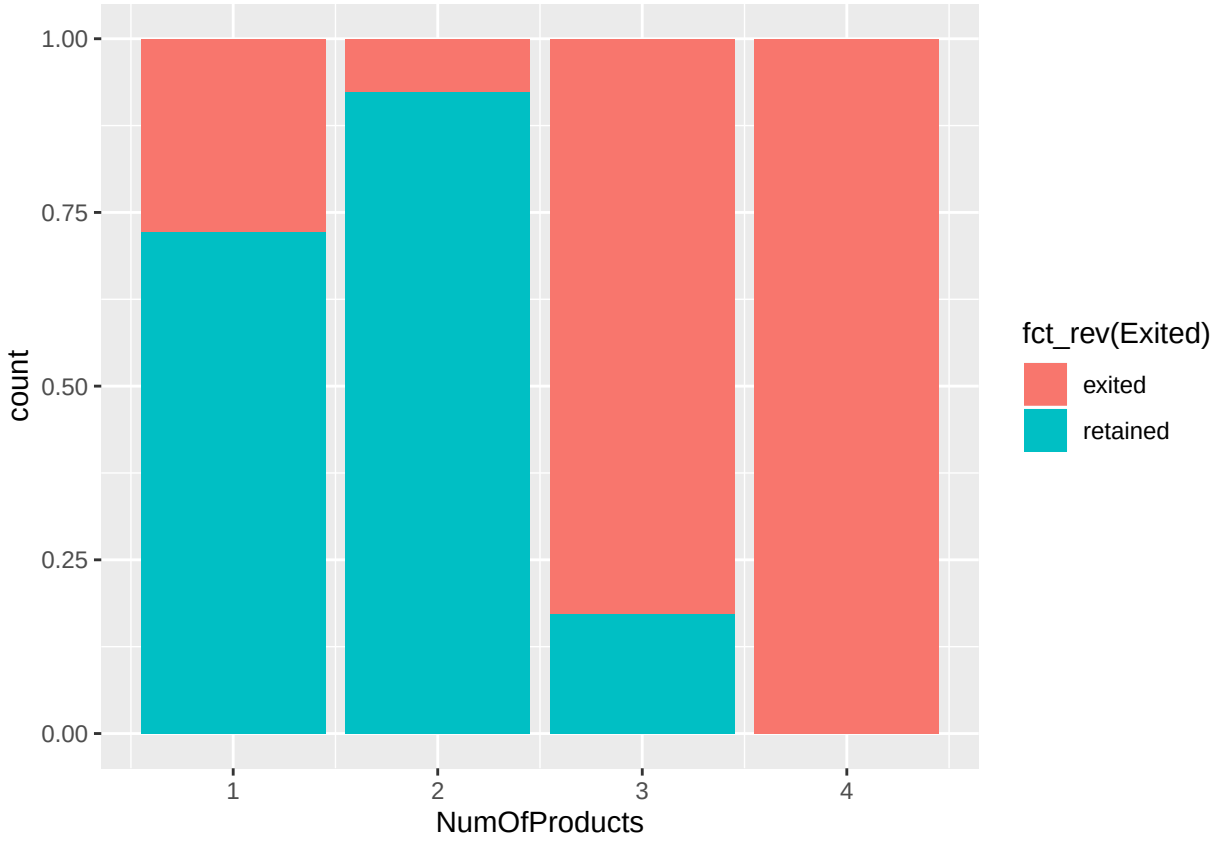

250000
200000
150000
100000
50000
0
Balance
retained
exited
Exited

Low balance has a better probability of being retained? We should be looking at box plots for age, tenure, and credit score.




800
700
600
500
400
CreditScore
retained
exited
Exited

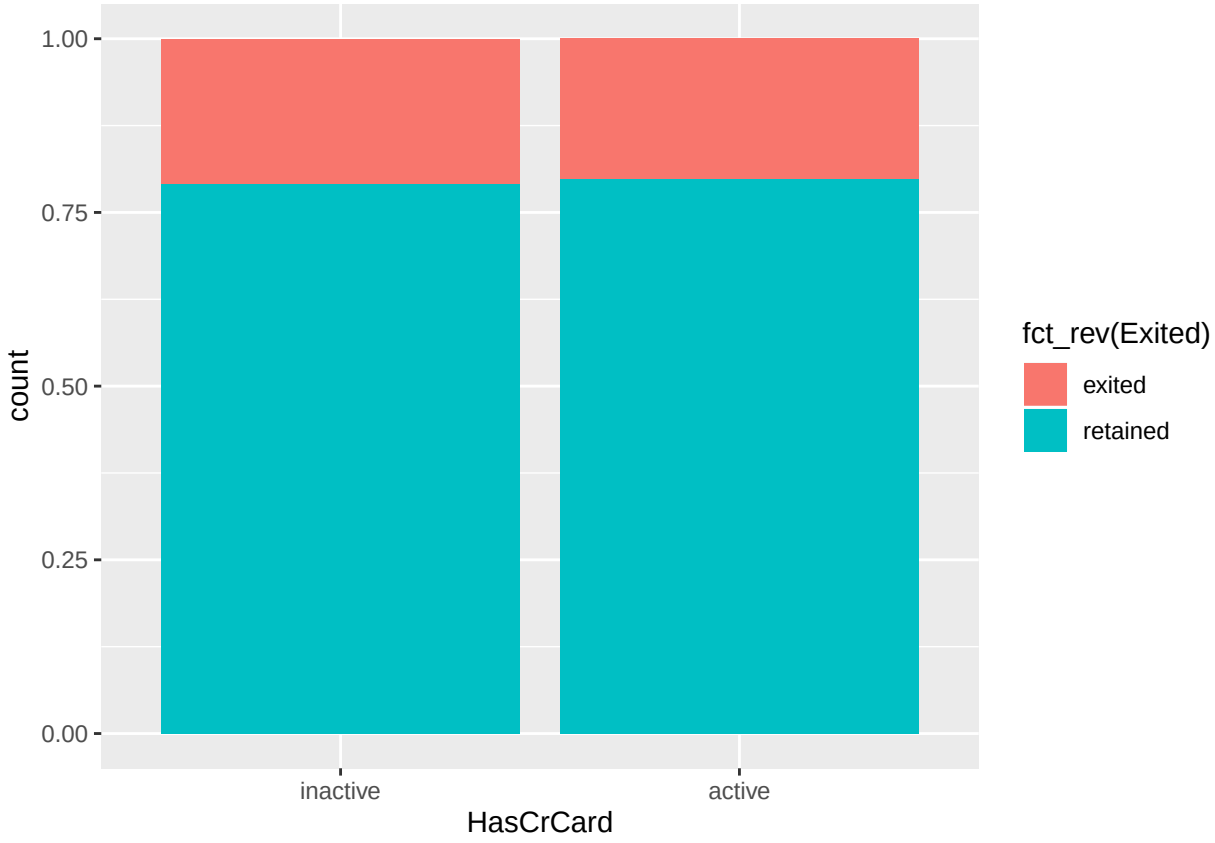
NUMBER OF PRODUCTS



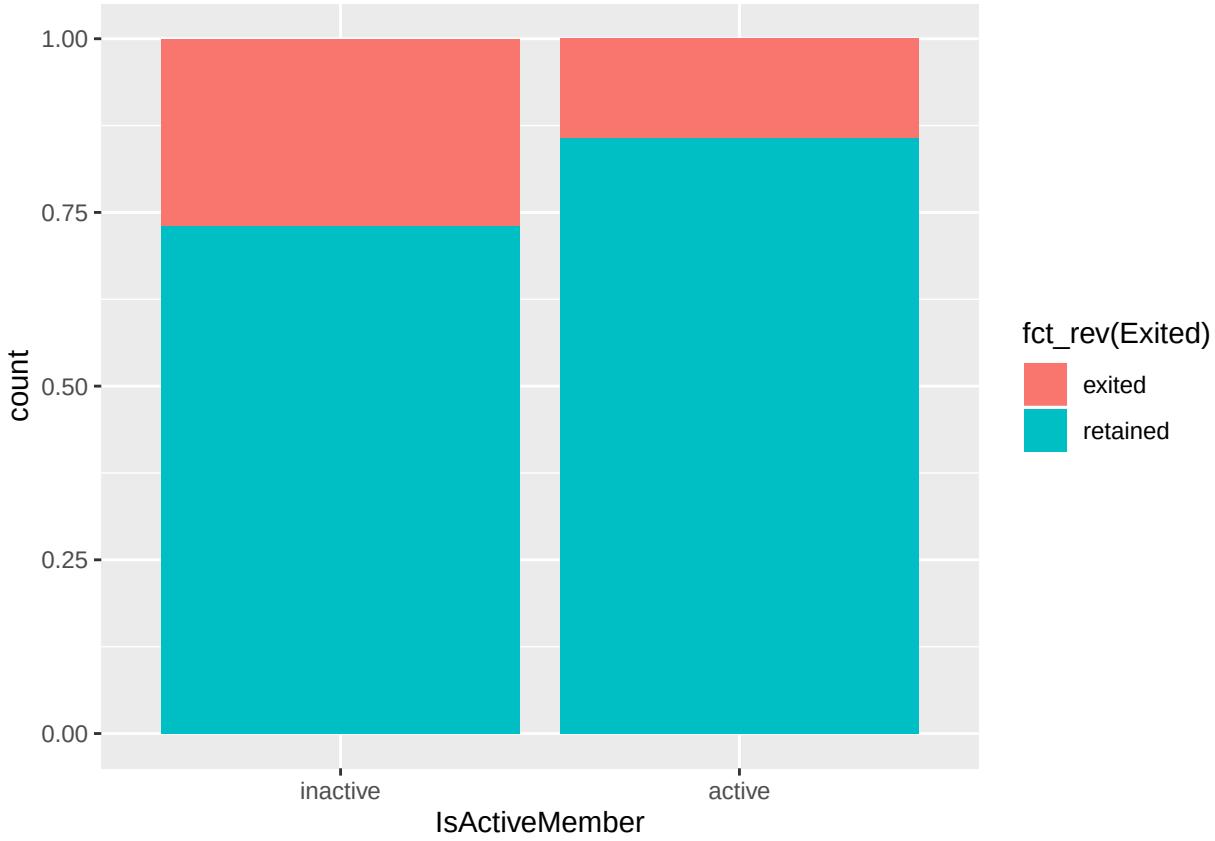

```
##
##    1    2    3    4
## 5084 4590  266   60
```

Groups are imbalanced, but still shows a high purity for some num of products based on this, it's more logical to treat NumOfProducts as a factor variable.

HAS CREDIT CARD

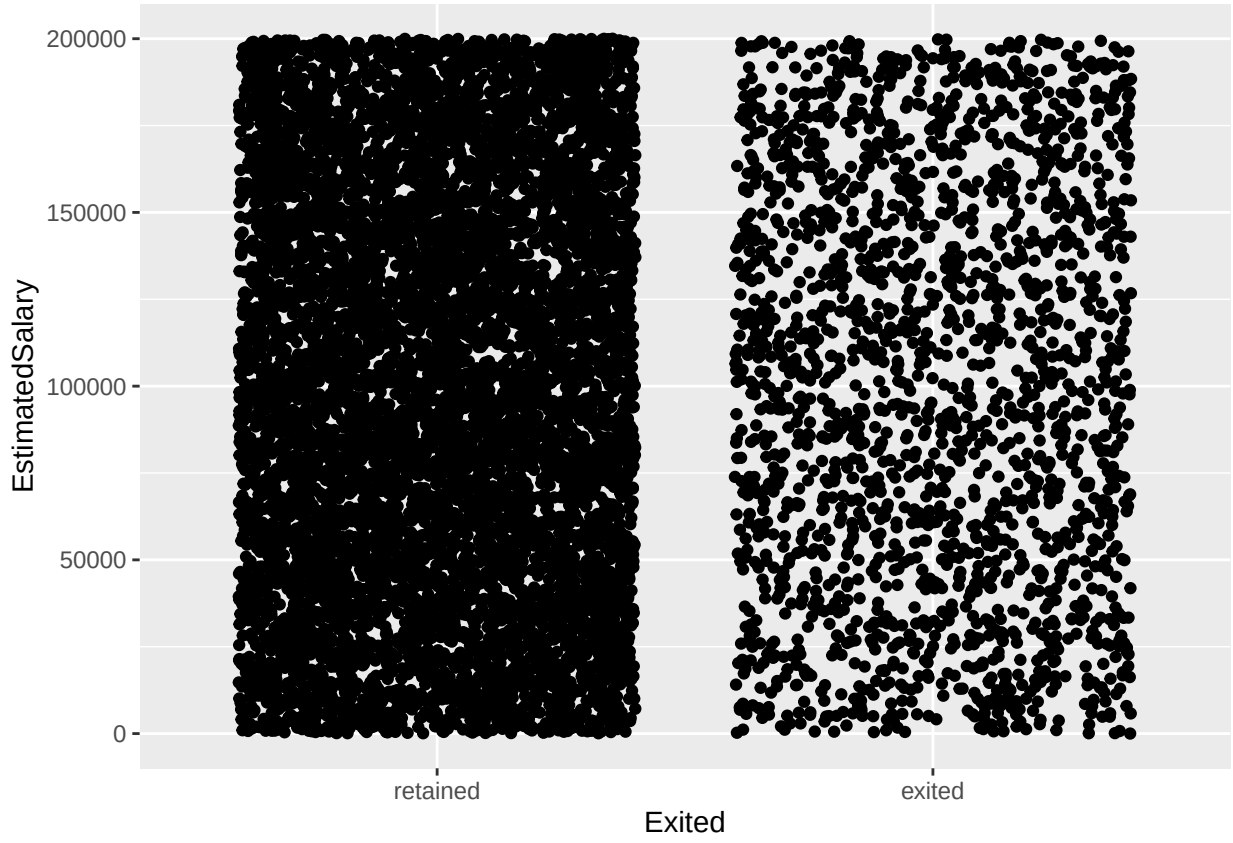


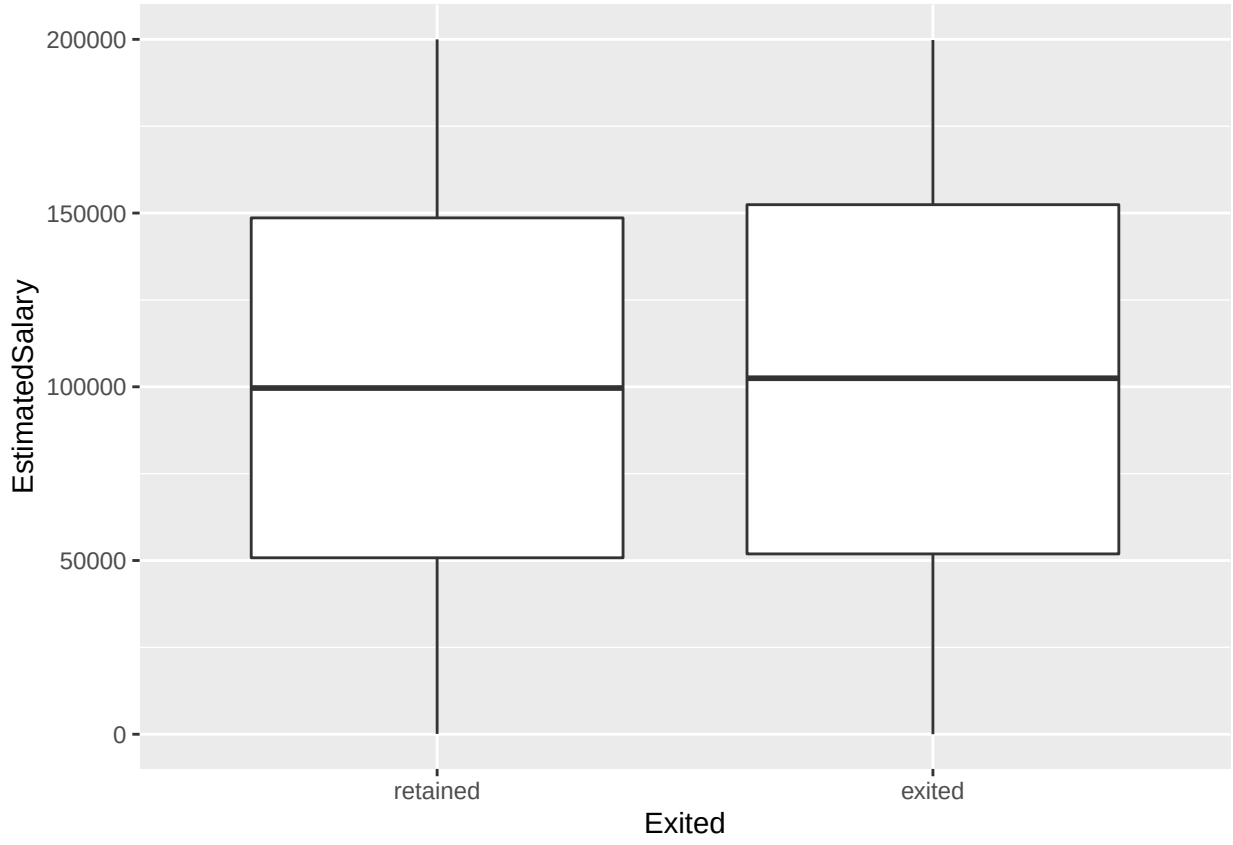
IS ACTIVE MEMBER



Does not provide a lot of insight or make much sense in the data set, will remove.

ESTIMATED SALARY





SURNAME