Exploratory Data Analysis

W.Lu

15/01/2020

Introduction

According to the customer survey about two different accessorie Brand , Elago and Belkin.

We according to the different factor (Age ,Salary, Car, Credit , Education Level ,Zipcode) to analyse the different factor which influence the consuming bahavior more .

As a result , we see findings as following.

1. In education factor, the customers who has level 2 and 3 consume more products .
2. In salary factor, the middle salary level (2500 ~ 10000) prefer brand Belkin more .The high salary level (> 10000) prefers Brand Elago.
3. In macro data ,high salary customer has lower credit .
4. In age group (< 30),the most customers buying Brand Elago, espicially for the customers has the Elevel 4. For Brank Belkin, could take attention for the following subgroup , which prefers belkin more than Elago.

   —> Age 25-30 (Elevel 3,Zipcode 1) (Elevel 3,Zipcode 7) (Elevel 3,Zipcode 8) (Elevel 2,Zipcode 3) (Elevel 2,Zipcode 7) (Elevel 2,Zipcode 8)

   —> Age 0-25 (Eleve2 3,Zipcode 6) (Eleve1 3,Zipcode 4) (Eleve1 3,Zipcode 6)
5. In age group (30-60),the most customers buying Brand Elago from Elevel 1 and 4.
6. In age group (>60),the most customers buying Brand Belkin.
7. In all customer group. The customers buying bit more Brand Elago products than Brand Belkin products.

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## corrplot 0.84 loaded
```

```
#----- check data types
summary(survey)
```

```
##          Salary          Age            Elevel           Car       
##  20000      : 129   Min.   :20.00   Min.   :1.000   Min.   : 1.00  
##  150000     : 116   1st Qu.:35.00   1st Qu.:2.000   1st Qu.: 5.00  
##  100014,3953:   1   Median :50.00   Median :2.000   Median :10.00  
##  100030,4649:   1   Mean   :49.81   Mean   :2.339   Mean   :10.43  
##  100050,7481:   1   3rd Qu.:65.00   3rd Qu.:3.000   3rd Qu.:16.00  
##  100051,1068:   1   Max.   :80.00   Max.   :4.000   Max.   :20.00  
##  (Other)    :9751                                                  
##     Zipcode          Credit        Brand     
##  Min.   :0.000   501,21 :   4   Belkin:4652  
##  1st Qu.:2.000   582,5  :   4   Elago :5348  
##  Median :4.000   658,66 :   4                
##  Mean   :4.037   716,15 :   4                
##  3rd Qu.:6.000   498,98 :   3                
##  Max.   :8.000   507,89 :   3                
##                  (Other):9978
```

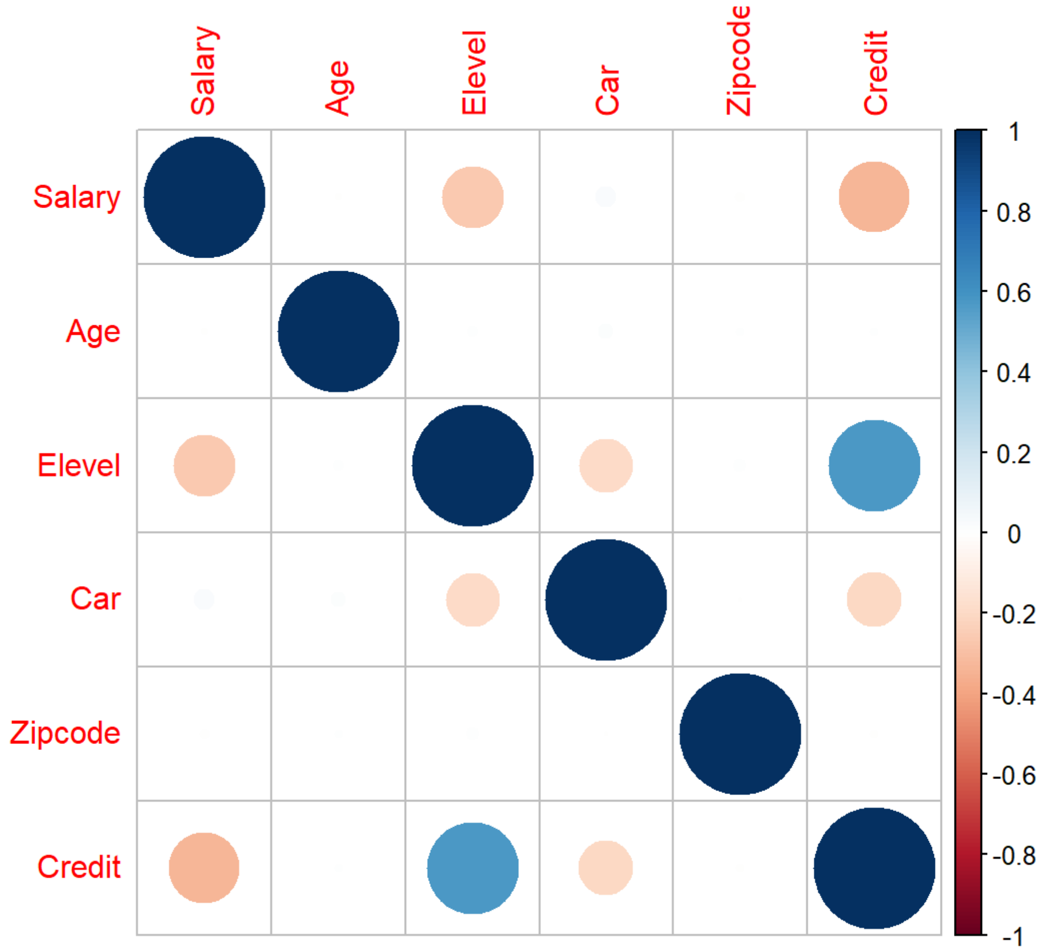
```
head(survey)
```

```
##        Salary Age Elevel Car Zipcode Credit  Brand
## 1 113770,6723  59      3  14       6 737,93 Belkin
## 2  139182,774  67      3  16       5 679,31  Elago
## 3 101966,6571  53      2   3       8 602,73 Belkin
## 4 71760,51794  26      2   8       5  653,9 Belkin
## 6  36716,5361  46      2   5       6 461,72  Elago
## 7 129555,8395  34      2  19       8 726,78  Elago
```

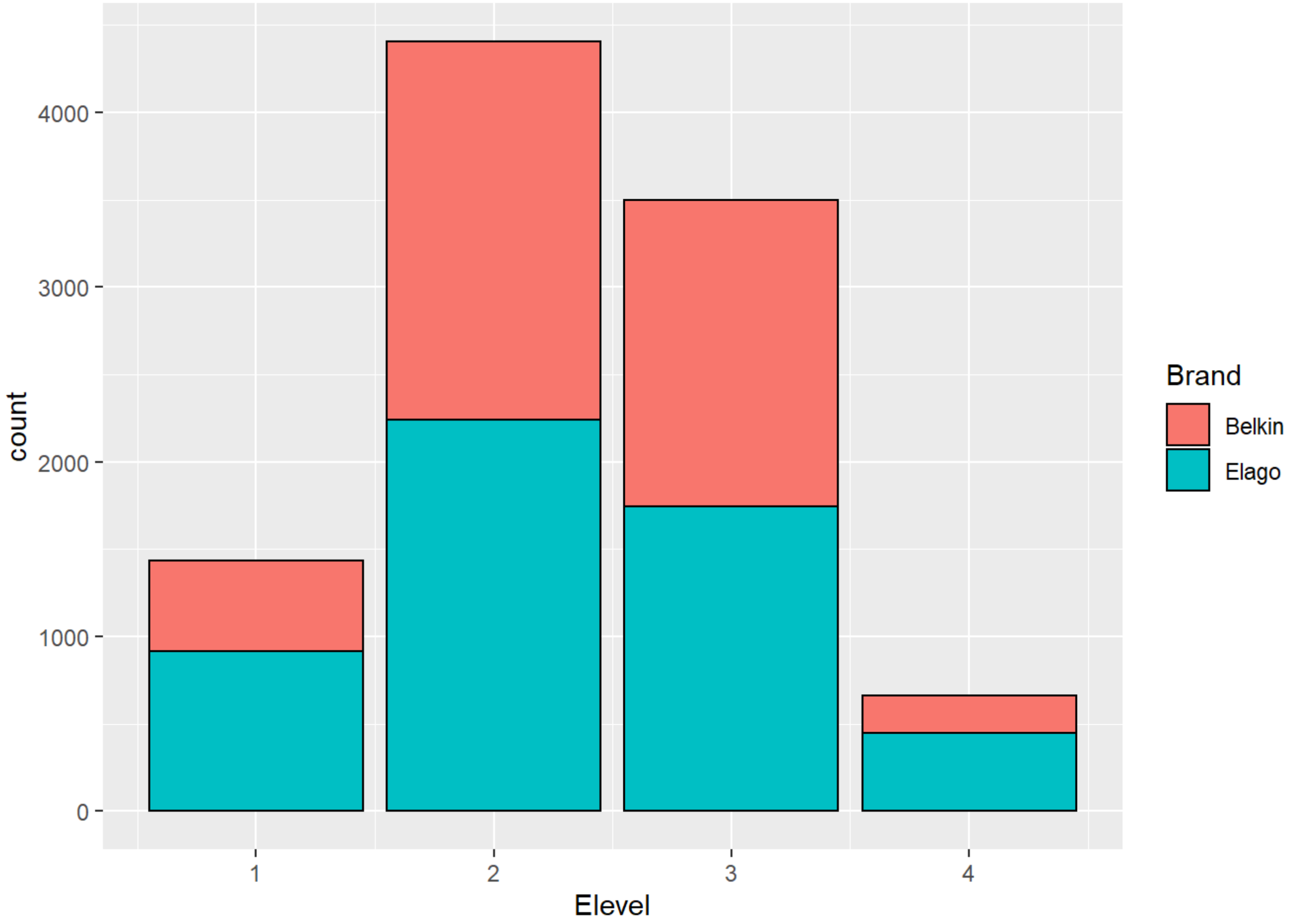
```
Youth<- survey[survey$Age < 30,]
Middel<- survey[survey$Age>30 & survey$Age < 60,]
Senior <- survey[survey$Age > 60,]

# correlation matrix
str(survey)
```

```
## 'data.frame':    10000 obs. of  7 variables:
##  $ Salary : Factor w/ 14757 levels "100010,30367054",..: 1592 4504 222 11448 7472 3403 138
94 13252 14553 13190 ...
##  $ Age    : int  59 67 53 26 46 34 71 26 67 48 ...
##  $ Elevel : int  3 3 2 2 2 2 2 2 3 2 ...
##  $ Car    : int  14 16 3 8 5 19 19 14 4 15 ...
##  $ Zipcode: int  6 5 8 5 6 8 4 5 0 3 ...
##  $ Credit : Factor w/ 11999 levels "423,71","426,94",..: 10152 7876 4338 6708 182 9765 920
3 5477 7230 8517 ...
##  $ Brand  : Factor w/ 2 levels "Belkin","Elago": 1 2 1 1 2 2 1 1 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int  5 8 10 15 17 19 22 23 24 25 ...
##   ..- attr(*, "names")= chr  "5" "8" "10" "15" ...
```

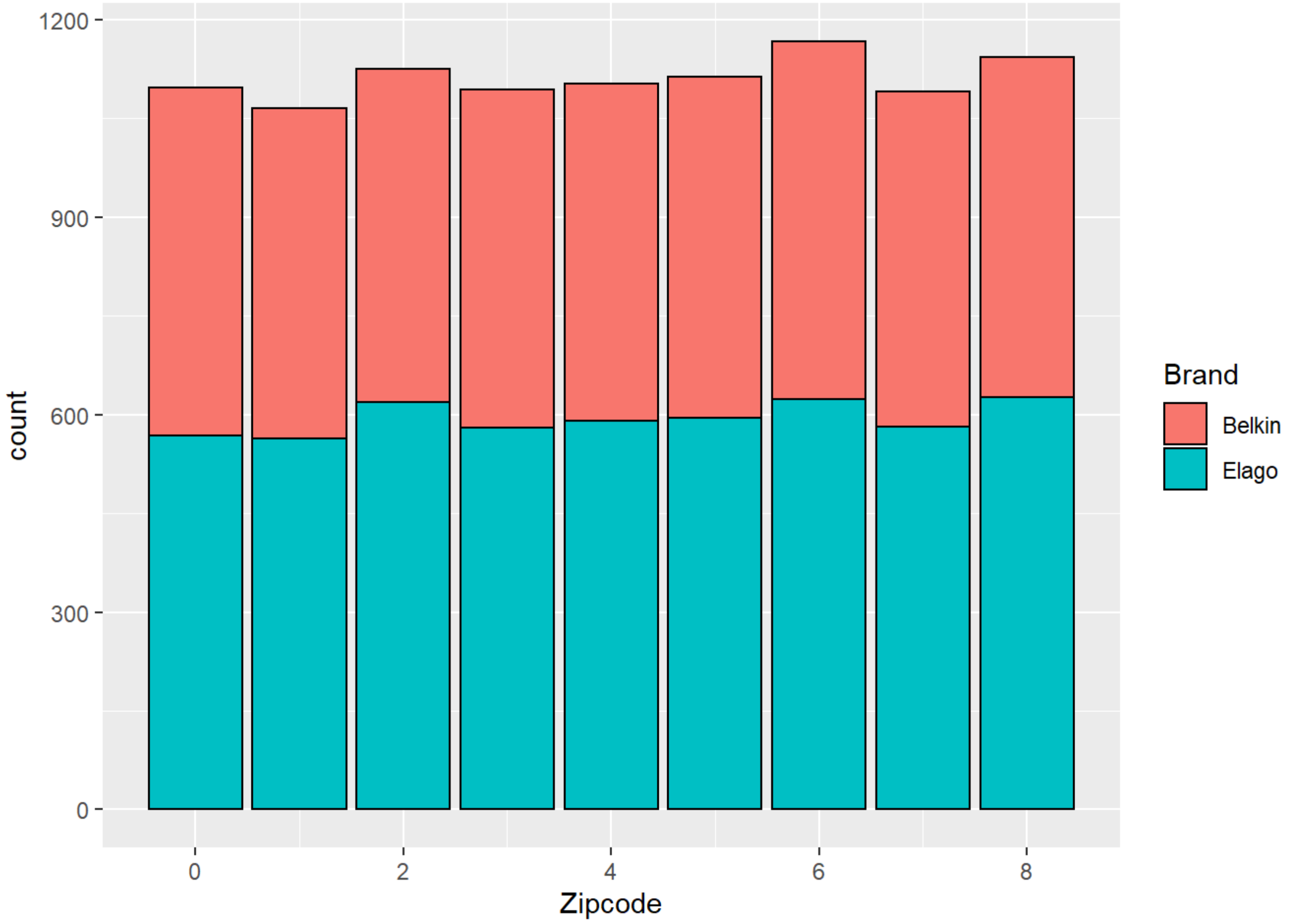
```
# change the factor to numeric
survey$Credit<-as.numeric(survey$Credit)
survey$Salary<-as.numeric(survey$Salary)
cm<-cor(survey[1:6])
corrplot(cm,)
```



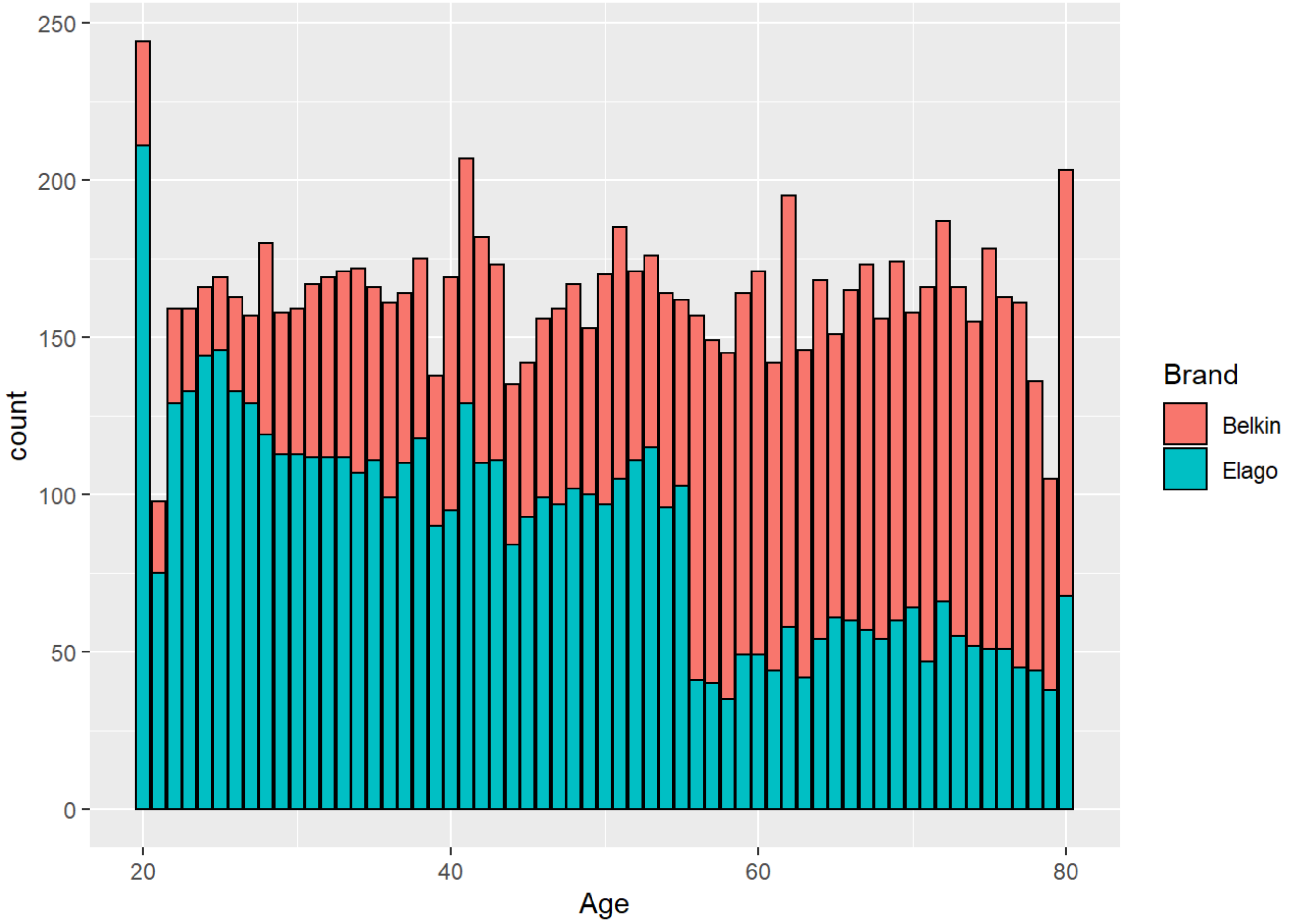
```
#----- visualization
#Chart 1 to 3 shows in different education level ,living area and Age , which brand are more popular .
d <- ggplot(data=survey,aes(x=Elevel))
d + geom_histogram(stat="count",binwidth = 10,aes(fill=Brand),colour="black")# 1 to 4 , 4 is higher eudcation
```



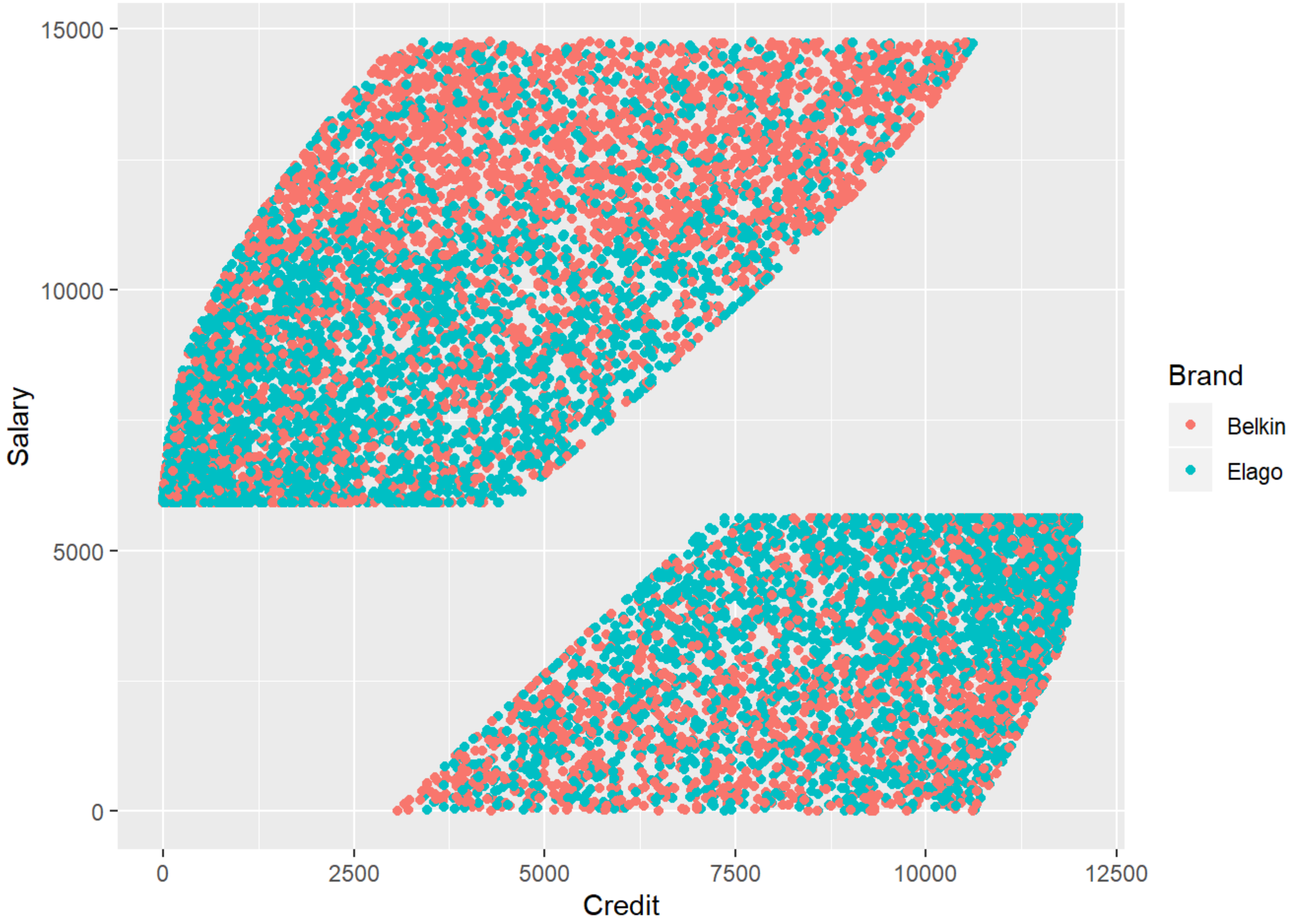
```
#----- chart 1 ,in education level 2 and 3 , customers has stronger consuming power.
e <- ggplot(data=survey,aes(x=Zipcode))
e + geom_histogram(stat="count",binwidth=10,aes(fill=Brand),colour="black")
```



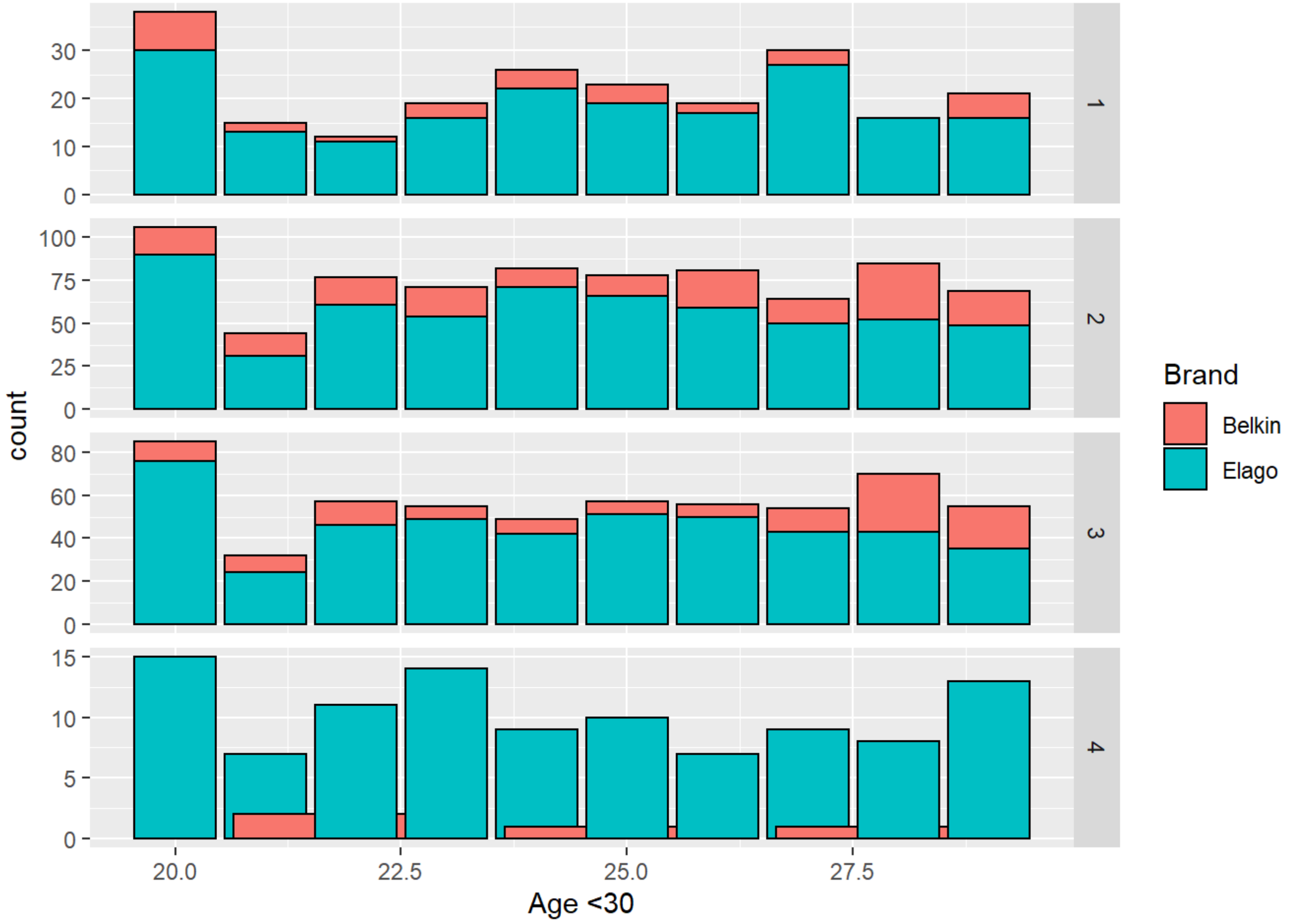
```
#----- chart 2 ,not obviously difference , in a macro way , elago has more market power .
f <- ggplot(data=survey,aes(x=Age))
f + geom_histogram(stat="count",binwidth=10,aes(fill=Brand),colour="black")
```



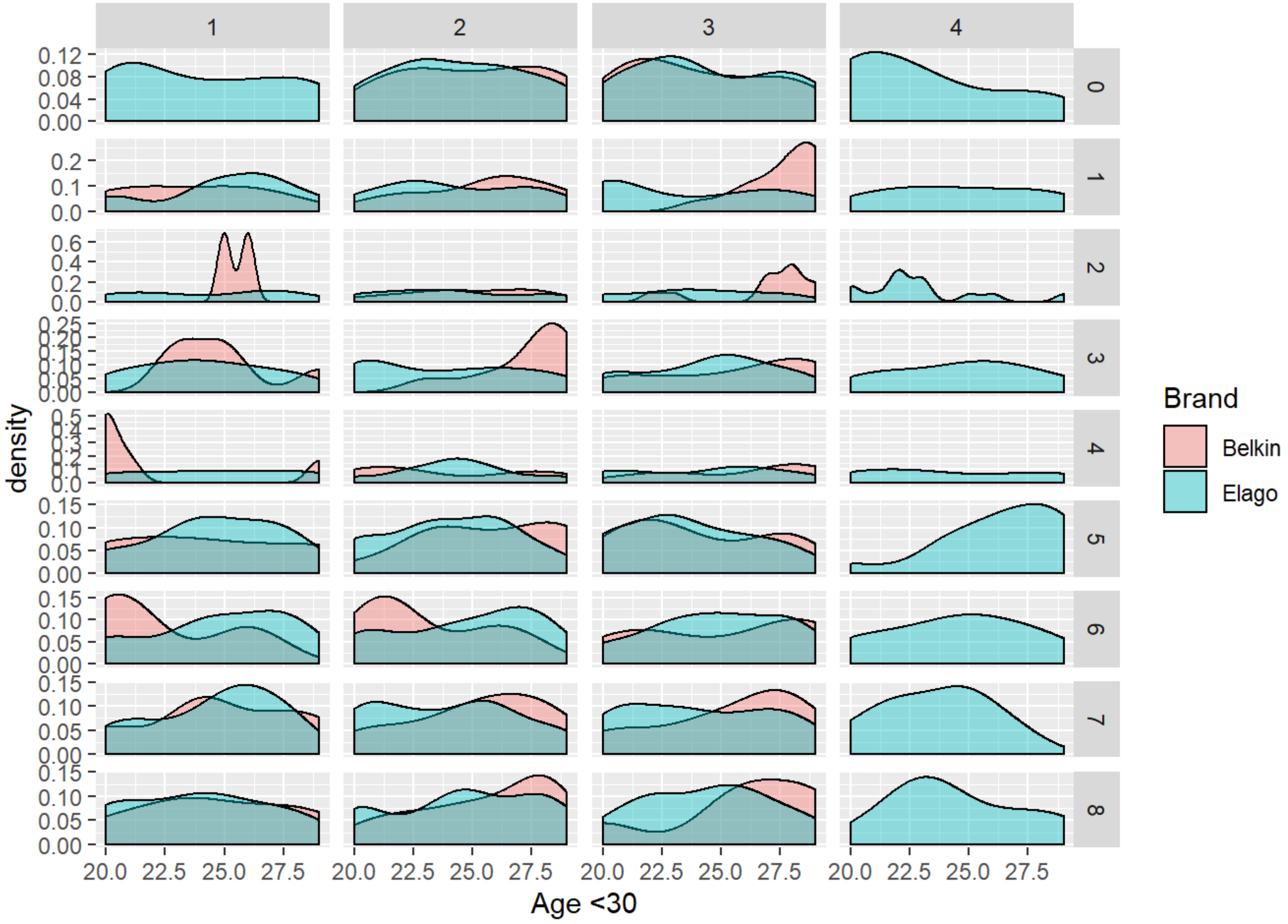
```
#----- chart 3, the customer group after 55 , prefers Belkin brand , and customer younger tha
n 55 ,prefer Elago.
g <- ggplot(data=survey,aes(x=Credit,y=Salary,colour=Brand))
g + geom_point()
```



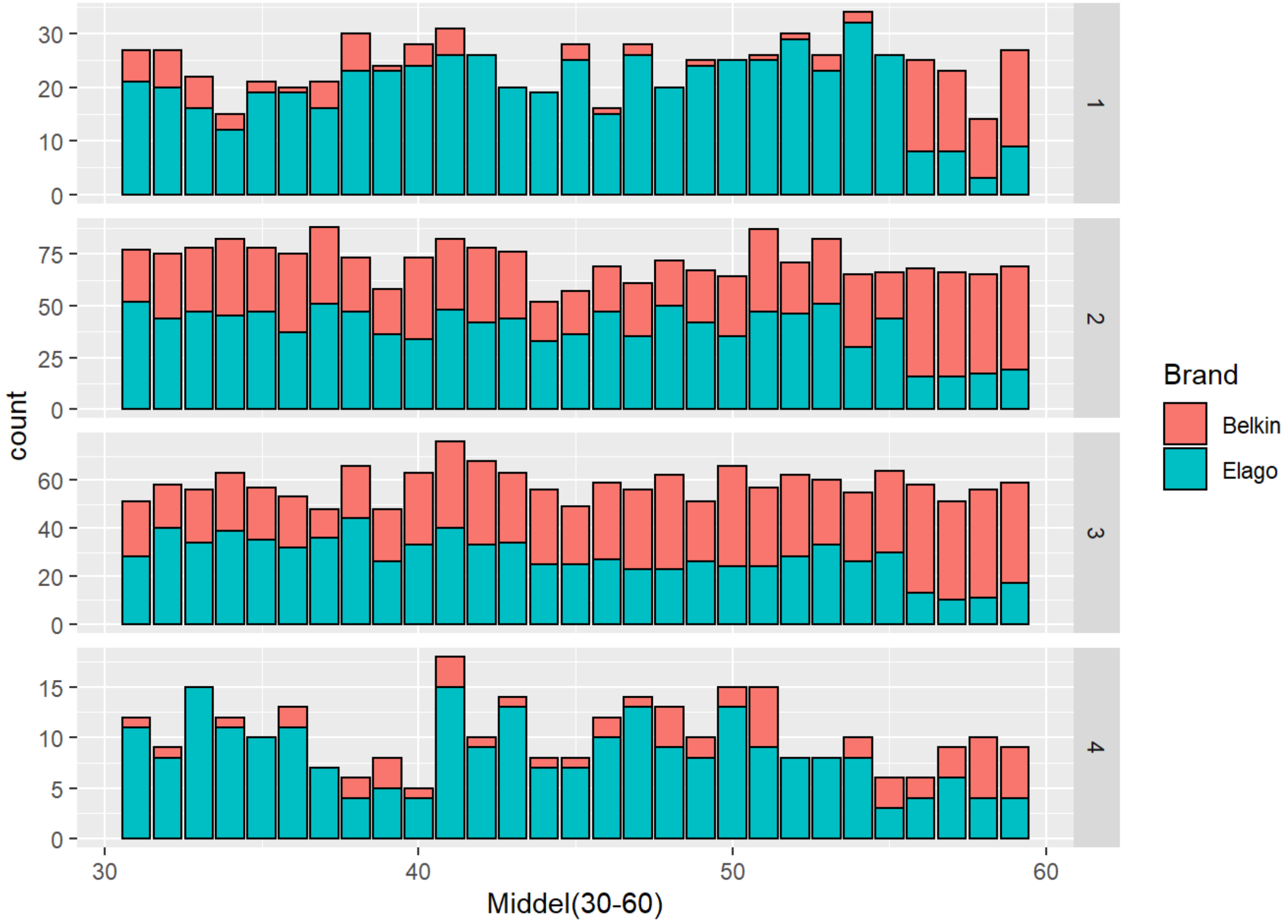
```
#----- chart 4, the customer has salary between 2500 and 10000, prefer Brand Belkin more , lo
wer credit , higher salary.
#------As age is a strong indictator for predicting customer behavior , so I have categorized
to 3 age groups ( Young, Middle, Senior )
a <- ggplot(data=Youth,aes(x=Age,colour=Brand))

a + geom_bar(aes(fill=Brand),colour="black")+facet_grid(Elevel~.,scale="free")+xlab("Age <30"
)
```



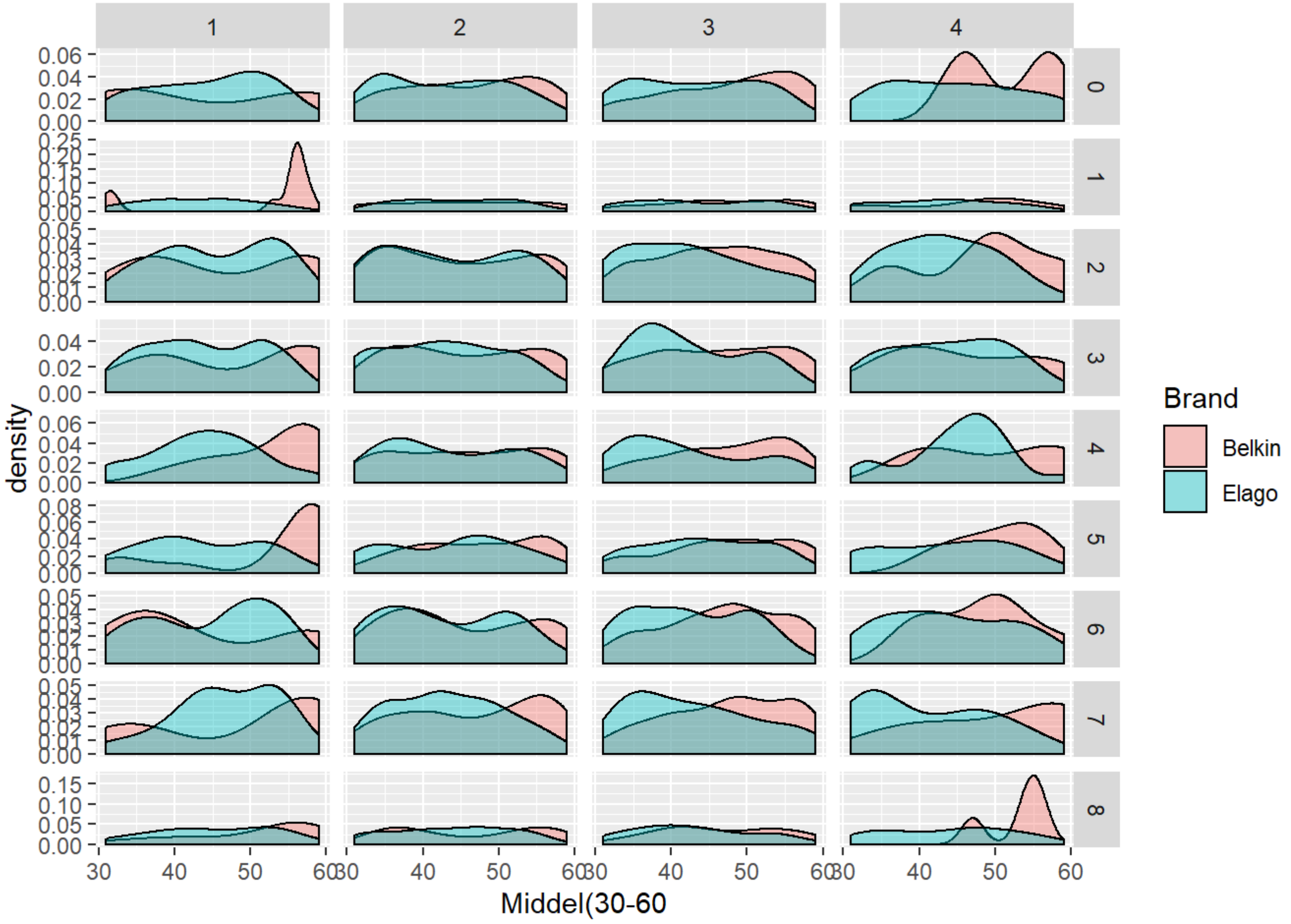
```
#----- Chart 5 shows in age group younger than 30,in defferent education group, the customer
 behavior relationship between age and brand.

a + geom_density(aes(fill=Brand),colour="black",alpha=0.4)+facet_grid(Zipcode~Elevel,scale="f
ree")+xlab("Age <30")
```



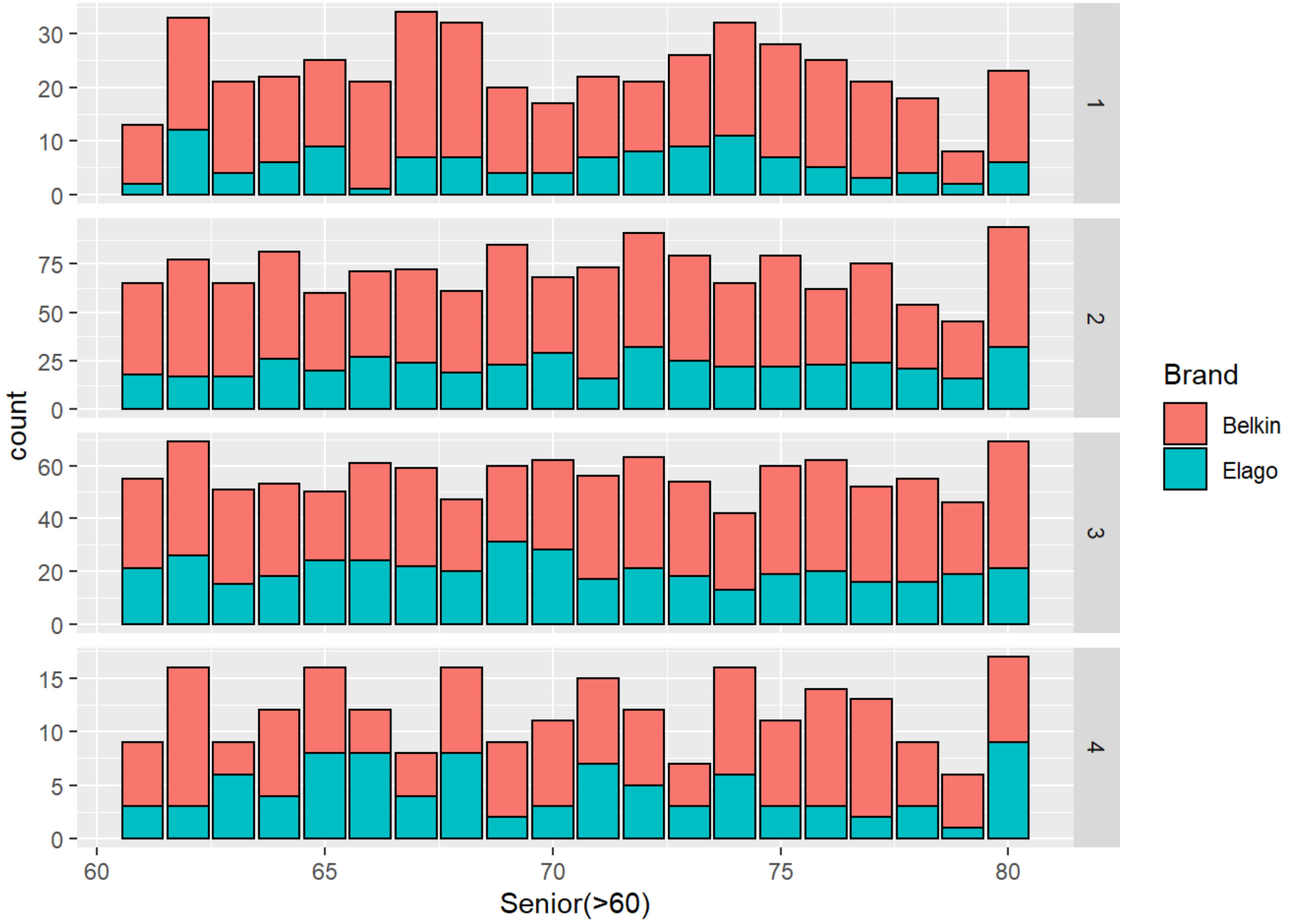
```
#----- Chart 6 shows in age group younger than 30 ,in defferent education group and different
Zipcode area , the customer behavior relationship with brand.
b <- ggplot(data=Middel,aes(x=Age,colour=Brand))
b + geom_bar(aes(fill=Brand),colour="black")+facet_grid(Elevel~.,scale="free")+xlab("Middel(3
0-60)")
```



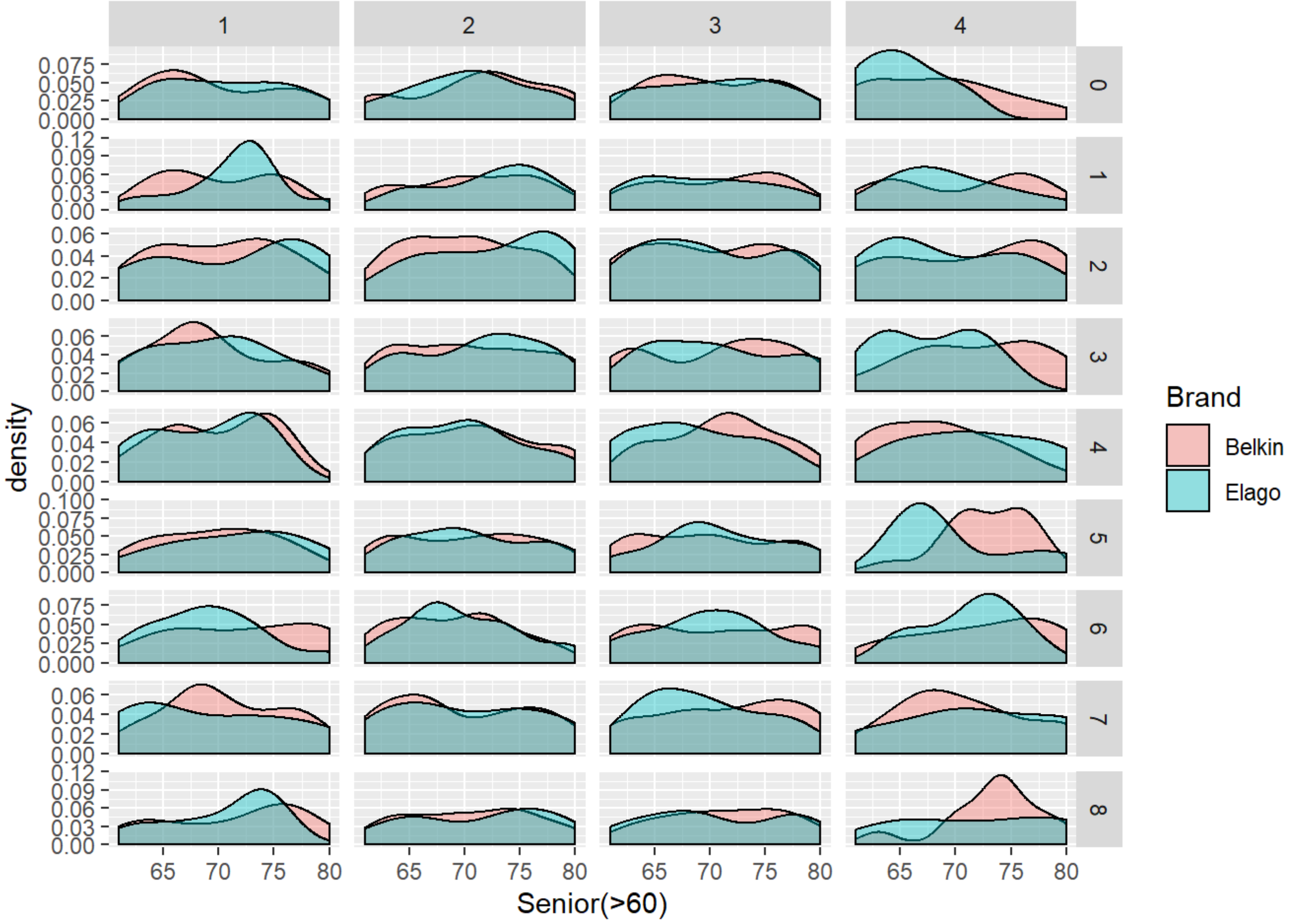
```
#----- Chart 7 shows in age group 31-60 ,in defferent education group, the customer behavior
 relationship between age and brand.
b + geom_density(aes(fill=Brand),colour="black",alpha=0.4)+facet_grid(Zipcode~Elevel,scale="f
ree")+xlab("Middel(30-60")
```



```
#----- Chart 8 shows in age group 31-60,in defferent education group and different Zipcode area , the customer behavior relationship with brand.
c <- ggplot(data=Senior,aes(x=Age,colour=Brand))
c + geom_bar(aes(fill=Brand),colour="black")+facet_grid(Elevel~.,scale="free")+xlab("Senior(>60)")
```



```
#----- Chart 9 shows in age group older than 60 ,in defferent education group, the customer behavior relationship between age and brand.
c + geom_density(aes(fill=Brand),colour="black",alpha=0.4)+facet_grid(Zipcode~Elevel,scale="free")+xlab("Senior(>60)")
```



```
#----- Chart 10 shows in age group older than 60 ,in defferent education group and different
 Zipcode area , the customer behavior relationship with brand.
```

Further questions

1. Why high salary customers prefer buying Elago products ?
2. Why people older than 60 prefer buying Belkin products?
3. Why in eudcation level 1 and 4 , customers prefer buying Elago products ?