ANA 515 Assignment 4

Fang Tham

2022-10-15

# Discuss the business problem/goal

Based on the mall customers dataset, I would like to segment the customers based on their age, gender, and interest to help with customized marketing for business owners in the mall.

# Retrieving data

The dataset is obtained from Kaggle: (<https://www.kaggle.com/datasets/shwetabh123/mall-customers?resource=download&select=Mall_Customers.csv>), titled “Mall\_Customers”, which contains information on customers such as customer id, gender, age, annual income, and spending score.

#I retrieve the dataset by setting a working directory   
setwd("/Users/FangTham/OneDrive - Hazel Technologies/McDaniel College/ANA515/Week 8")  
malldata <- read.csv("Mall\_Customers.csv")

I printed the first ten rows of data to ensure the dataset is imported

head (malldata)

## CustomerID Genre Age Annual.Income..k.. Spending.Score..1.100.  
## 1 1 Male 19 15 39  
## 2 2 Male 21 15 81  
## 3 3 Female 20 16 6  
## 4 4 Female 23 16 77  
## 5 5 Female 31 17 40  
## 6 6 Female 22 17 76

# Cleaning up the data set

I realized that the column name “gender” is mispelled, so i renamed the column, and cleaned up the names of the other columns

colnames(malldata)[2] = "Gender"  
colnames(malldata)[4] = "AnnualIncome"  
colnames(malldata)[5] = "SpendingScore"

I inspect the characteristics of my dataset

This data set has 200 rows and 5 columns. The names of the columns and a brief description of each are in the table below:

## Names Description  
## 1 CustomerID Customer number ID  
## 2 Gender Gender of customer  
## 3 Age Age of customer  
## 4 AnnualIncome Annual income of customer  
## 5 SpendingScore Assigned spending score of customer from 1-100

# Summarizing the data set

#This is to show a summary of the dataset   
summary(malldata)

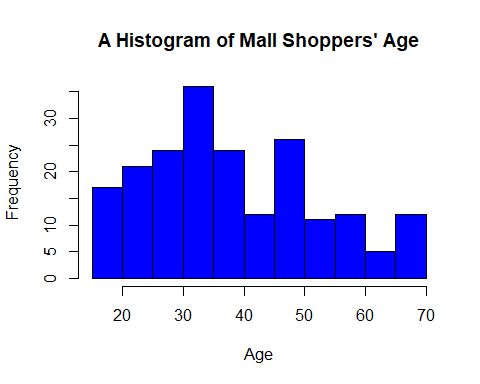
## CustomerID Gender Age AnnualIncome   
## Min. : 1.00 Length:200 Min. :18.00 Min. : 15.00   
## 1st Qu.: 50.75 Class :character 1st Qu.:28.75 1st Qu.: 41.50   
## Median :100.50 Mode :character Median :36.00 Median : 61.50   
## Mean :100.50 Mean :38.85 Mean : 60.56   
## 3rd Qu.:150.25 3rd Qu.:49.00 3rd Qu.: 78.00   
## Max. :200.00 Max. :70.00 Max. :137.00   
## SpendingScore   
## Min. : 1.00   
## 1st Qu.:34.75   
## Median :50.00   
## Mean :50.20   
## 3rd Qu.:73.00   
## Max. :99.00

table(malldata$Gender)/length(malldata$Gender)

##   
## Female Male   
## 0.56 0.44

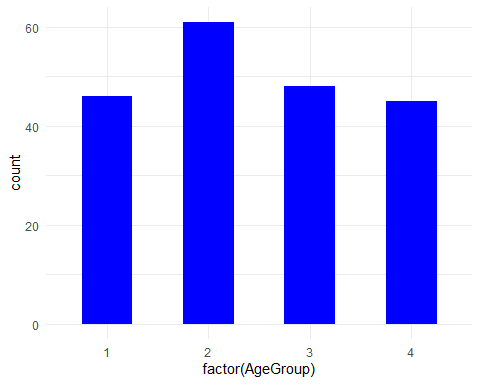
Based on the dataset, there are 56% of female, and 44% of male shoppers. The youngest age of a customer is 18, and the oldest age of a customer is 70, with a mean of 38, and median of 36. This means that the data is just slightly right skewed. We also saw that the spending score ranged from 1-00, with a mean of 50.20 and median of 50. The annual income ranged from 15k-137k. To better understand the distribution, I will plot a histogram for the age group.

hist(malldata$Age, col="blue",border="black", xlab="Age", main="A Histogram of Mall Shoppers' Age")



I then split the age groups into 4 distint age groups based on age

Agegrouped <- malldata %>%   
 mutate(AgeGroup = case\_when(Age>=18 & Age <28 ~'1',  
 Age>=28 & Age <38 ~ '2',  
 Age>=38 & Age<50 ~ '3',  
 Age>= 50 ~ '4'))  
  
ggplot(Agegrouped, aes(x=factor(AgeGroup)))+  
 geom\_bar(stat="count", width=0.5, fill="blue")+  
 theme\_minimal()



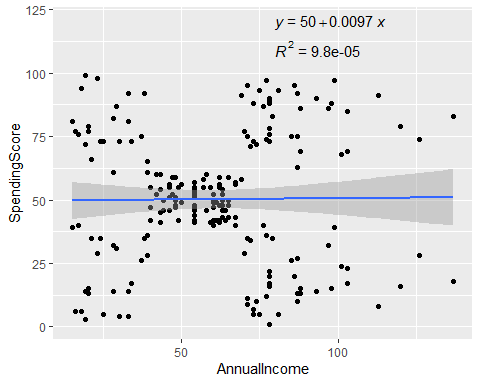
We can see the age group “2” builds the highest proportion of shoppers, and age group 1, 3, and 4 are pretty equal in proportion. I wanted to see if there’s a correlation between the age group and their annual income, so I perform a cross tab analysis on the groups and mean annual income.

aggregate(x=Agegrouped$AnnualIncome,  
 by=list(Agegrouped$AgeGroup),   
 FUN=mean)

## Group.1 x  
## 1 1 48.50000  
## 2 2 70.08197  
## 3 3 65.81250  
## 4 4 54.37778

I found that age group “2” of 28 to 38 has the highest annual income of 70k, where as group “1” of 18 to 28 years old has the lowest income. This make sense as younger shoppers are probably still in college and not pursuing a full time job yet. The older shoppers are also probably retired. I wanted to see if there is a correlation between the annual income and the spending score.

#Scatter plot of spending score vs annual income  
incomespending <- ggplot(Agegrouped, aes(AnnualIncome, SpendingScore)) +   
 geom\_point()  
  
graphplot <- incomespending +   
 stat\_smooth(method = "lm",formula = y ~ x,geom = "smooth") +  
 stat\_regline\_equation(label.x=80, label.y=120) +  
 stat\_cor(aes(label=..rr.label..), label.x=80, label.y=110)  
graphplot



The R squared value is 9.8E-5, which means that the correlation between Spending Score and Annual Income is low. I had expected to get a positive correlation. However, it is interesting to see that the concentration of spending score is tighter (around a score of 30-60) for shoppers with annual income around 48k-60k compared to all other shoppers.

I also wanted to see if gender and age group plays a role in spending score

aggregate(x=Agegrouped$SpendingScore,  
 by=list(Agegrouped$Gender, Agegrouped$AgeGroup),   
 FUN=mean)

## Group.1 Group.2 x  
## 1 Female 1 58.75000  
## 2 Male 1 52.22727  
## 3 Female 2 63.15789  
## 4 Male 2 59.00000  
## 5 Female 3 37.64286  
## 6 Male 3 47.10000  
## 7 Female 4 41.22727  
## 8 Male 4 35.69565

From this, we can see that there is a general trend where females have higher spending scores than males. Moreover, age group 2 (28-38) also had the highest mean spending score for both male and females.

# Conclusion

Overall, we can conclude that: 1. There is no direct correlation between annual income and spending score 2. Females have higher spending score than males across all age groups 3. Age group 2 (28-38 years old) have the highest spending score