Assignment 4: Mall Customers Segmentation

Allison Chaang

4/29/2021

# Introduction to the Dataset

The business problem that I want to solve is to understand the demographics of mall shoppers and to empower businesses to cater their products to their specific target markets. With this dataset of 200 consumers, we can help businesses narrow down which groups of consumers have higher spending power and how to segregate them for customized marketing and advertising.

# Importing and Cleaning the dataset

The dataset that I am using is linked [here](https://www.kaggle.com/shwetabh123/mall-customers?select=Mall_Customers.csv), titled *“The Dataset for Frequent Mall Customers”*, which contains information on customers such as their gender, customer id, age, annual income, and spending score.

I downloaded the dataset into my desktop and imported it from my working directory.

setwd("/Users/allisonchaang/Desktop/")  
getwd()

## [1] "/Users/allisonchaang/Desktop"

mallcustomersdata <- read.csv("mallcustomers.csv")

Next, I wanted to see whether all the data was imported into R so I printed the first 10 rows of the dataset.

head(mallcustomersdata,10)

## 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  
## 7 7 Female 35 18 6  
## 8 8 Female 23 18 94  
## 9 9 Male 64 19 3  
## 10 10 Female 30 19 72

Based on the code below, we see that there are 5 columns of data for the customer ID, gender, age, annual income and spending score. Let’s rename some of the columns to clean any spelling errors and ensure that it accurately explains the data.

ncol(mallcustomersdata)

## [1] 5

colnames(mallcustomersdata)

## [1] "CustomerID" "Genre" "Age"   
## [4] "Annual.Income..k.." "Spending.Score..1.100."

library(dplyr)

##   
## 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

newtitles <- rename(mallcustomersdata,Gender=Genre, AnnualIncome=Annual.Income..k..,SpendingScore=Spending.Score..1.100.)

I want to make make sure the column names are updated

colnames(newtitles)

## [1] "CustomerID" "Gender" "Age" "AnnualIncome"   
## [5] "SpendingScore"

Next, I want to make make sure the annual income is reflected in 1000s

library(stringr)  
newtitles$AnnualIncome=str\_pad(newtitles$AnnualIncome,width=5,side="right",pad="0")

head(newtitles)

## CustomerID Gender Age AnnualIncome SpendingScore  
## 1 1 Male 19 15000 39  
## 2 2 Male 21 15000 81  
## 3 3 Female 20 16000 6  
## 4 4 Female 23 16000 77  
## 5 5 Female 31 17000 40  
## 6 6 Female 22 17000 76

# Exploring the dataset

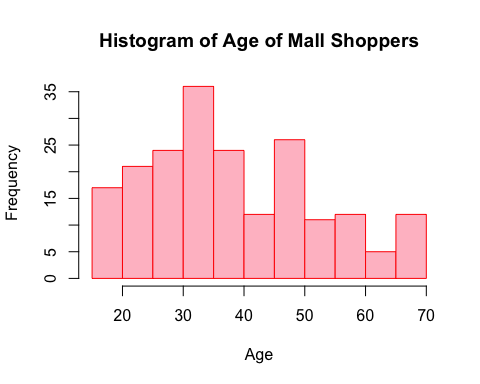
We see that there are 200 unique customer IDS, and the customers range from 18 years old to 70 years old. On average though, the customers are around 39 years old so most of them are middle-aged consumers.

summary(newtitles)

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

Given that there is such a large variation in terms of ages of consumers, as shown in the histogram below, I want to group them by 4 age groups: 1 (15-29), 2 (30-44), 3 (45-59) and 4(60 and above) to better understand which age group shops the most by average.

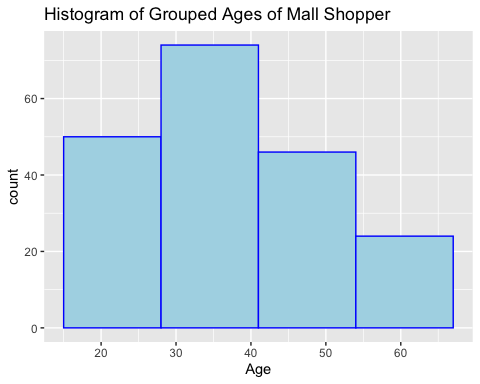
hist(newtitles$Age, col="pink",border="red", xlab="Age", main="Histogram of Age of Mall Shoppers")



library(dplyr)  
grouped <- newtitles %>% mutate(AgeGroup = case\_when(Age>=15 & Age <28 ~'1',  
 Age>=28 & Age <41 ~ '2',  
 Age>=41 & Age<54 ~ '3',  
 Age>= 54 ~ '4'))

Next, I want to re-create the histogram of consumer age but with bins by *Age* so that it can be visually easier to understand as well as easier to analyze the data.

library(ggplot2)  
ggplot(data=grouped, aes(Age))+geom\_histogram(breaks=seq(15,70,by=13),fill="light blue",col="blue" ) +labs(title="Histogram of Grouped Ages of Mall Shopper")

 From the output below, we can see that the spending score is highest among groups 1 and 2 which are those between the ages of 15-41 in aggregate. The rate slowly decreases as the consumer group gets older. However, the gap between the 4 groups are not too large, they only vary by ~15 of spending score.

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

## Group.1 x  
## 1 1 55.63043  
## 2 2 61.51316  
## 3 3 33.90909  
## 4 4 38.64706

Next, I want to dig a little deeper into the age groups and their annual income. Since groups 1 and 2 have the highest spending score, I wonder if it correlates with higher annual income. However, I noticed that the AnnualIncome column is in characters format so I changed it to integer using the as.integer function

str(grouped)

## 'data.frame': 200 obs. of 6 variables:  
## $ CustomerID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Gender : chr "Male" "Male" "Female" "Female" ...  
## $ Age : int 19 21 20 23 31 22 35 23 64 30 ...  
## $ AnnualIncome : chr "15000" "15000" "16000" "16000" ...  
## $ SpendingScore: int 39 81 6 77 40 76 6 94 3 72 ...  
## $ AgeGroup : chr "1" "1" "1" "1" ...

grouped$AnnualIncome=as.integer(grouped$AnnualIncome)  
str(grouped)

## 'data.frame': 200 obs. of 6 variables:  
## $ CustomerID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Gender : chr "Male" "Male" "Female" "Female" ...  
## $ Age : int 19 21 20 23 31 22 35 23 64 30 ...  
## $ AnnualIncome : int 15000 15000 16000 16000 17000 17000 18000 18000 19000 19000 ...  
## $ SpendingScore: int 39 81 6 77 40 76 6 94 3 72 ...  
## $ AgeGroup : chr "1" "1" "1" "1" ...

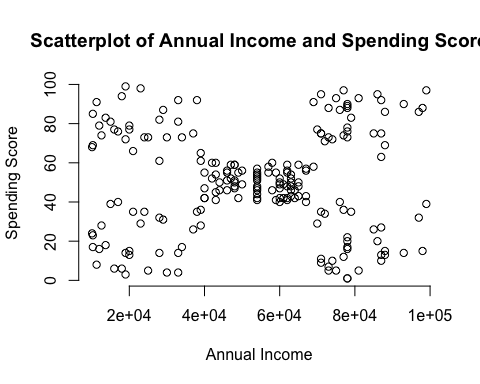
Now that the AnnualIncome is in integer format, I can proceed to see the average annual income of consumers across the 4 age groups. From the data below, we see that group 2 (28-41 years old) has the highest average income of $56,178. This makes sense that they also have a high spending score. While consumers in Group 1 (below 28) also have a high spending score, I think it also makes sense that they have lower average annual incomes compared to Groups 2 and 3 because they are younger thus they do not have the upper hand to aim for higher positions / compensation packages based on seniority. Perhaps we can conclude that Groups 3 to 4 have sustainable incomes but frequent these shopping malls less compared to younger consumers.

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

## Group.1 x  
## 1 1 48500.00  
## 2 2 56178.95  
## 3 3 54634.09  
## 4 4 51826.47

I also wanted to see if there was a correlation between spending score and annual income. Naturally I assumed that a high income would lead to higher spending score, vice versa. However, I am surprised by the scatterplot below that does not show a positive correlation between the two. Most of the data aggregates in the middle of annual income and spending score, which renders this correlation inconclusive.

x<- grouped$AnnualIncome  
y <- grouped$SpendingScore  
plot(x,y,main="Scatterplot of Annual Income and Spending Score", xlab="Annual Income",ylab="Spending Score", frame=FALSE)



Also, I’d like to also point out that of the 200 consumers we surveyed in this data, 112 of them are female and 88 of them are males. I have a gut feeling that gender might play a role in the spending patterns of consumers so I want to incorporate their gender into analysis.

aggregate(x=grouped$CustomerID,  
 by=list(grouped$Gender),   
 FUN=length)

## Group.1 x  
## 1 Female 112  
## 2 Male 88

Across all both genders, we can see that females across all ages fare higher than males in terms of spending scores. To piece it all together, we can conclude that younger, female shoppers tend to spend the most thus they would be the most lucrative market segment to target for businesses in the mall.

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

## Group.1 Group.2 x  
## 1 1 Female 58.75000  
## 2 2 Female 61.73333  
## 3 3 Female 35.17857  
## 4 4 Female 39.86667  
## 5 1 Male 52.22727  
## 6 2 Male 61.19355  
## 7 3 Male 31.68750  
## 8 4 Male 37.68421

Additionally, while females across all age groups have lower annual incomes than their male counterparts, females are still showing higher spending scores compared to men. This shows that there must be other extrinsic or intrinsic motivation for females to spend, there is no positive correlation between annual income and spending score. A higher income does not necessarily lead to or correlate with a more spenidng habits.

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

## Group.1 Group.2 x  
## 1 1 Female 45958.33  
## 2 2 Female 55271.11  
## 3 3 Female 50603.57  
## 4 4 Female 49073.33  
## 5 1 Male 51272.73  
## 6 2 Male 57496.77  
## 7 3 Male 61687.50  
## 8 4 Male 54000.00