ITCS 1880: Final Project

Michael Beebe

Loading in dependencies

library(tidyverse)  
library(glue)  
library(ggplot2)  
library(hrbrthemes)  
library(psych)  
library(graphics)  
library(dplyr)

Loading in dataset to be analyzed

customer\_data <- read.csv("Data/Mall\_Customers.csv")

# Exploratory Data Analysis

The number of rows and the number of columns in the data set.

nrow(customer\_data)

## [1] 200

ncol(customer\_data)

## [1] 5

The names of the columns in the data set

colnames(customer\_data)

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

The top 10 rows of the data set

head(customer\_data, 10)

## CustomerID Gender 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

The class of the dataset and each column

class(customer\_data)

## [1] "data.frame"

lapply(customer\_data, class)

## $CustomerID  
## [1] "integer"  
##   
## $Gender  
## [1] "character"  
##   
## $Age  
## [1] "integer"  
##   
## $Annual.Income..k..  
## [1] "integer"  
##   
## $Spending.Score..1.100.  
## [1] "integer"

Summarize the data in the Mall Customers data set. Be sure to include: Min, Max, Mean, Median, Quartiles and Standard Deviation:

### Age

Min

glue("Min Age: {min(customer\_data$Age)}")

## Min Age: 18

Max

glue("Max Age: {max(customer\_data$Age)}")

## Max Age: 70

Mean

glue("Mean Age: {mean(customer\_data$Age)}")

## Mean Age: 38.85

Median

glue("Median Age: {median(customer\_data$Age)}")

## Median Age: 36

Quartiles

quantile(customer\_data$Age)

## 0% 25% 50% 75% 100%   
## 18.00 28.75 36.00 49.00 70.00

Standard Deviation

glue("Standard Deviation of Age: {sd(customer\_data$Age)}")

## Standard Deviation of Age: 13.9690073315589

### Annual Income ($1,000’s)

Min

glue("Min Annual Income in $1,000's: {min(customer\_data$Annual.Income..k..)}")

## Min Annual Income in $1,000's: 15

Max

glue("Max Annual Income in $1,000's: {max(customer\_data$Annual.Income..k..)}")

## Max Annual Income in $1,000's: 137

Mean

glue("Mean Annual Income in $1,000's: {mean(customer\_data$Annual.Income..k..)}")

## Mean Annual Income in $1,000's: 60.56

Median

glue("Median Annual Income in $1,000's: {median(customer\_data$Annual.Income..k..)}")

## Median Annual Income in $1,000's: 61.5

Quartiles

quantile(customer\_data$Annual.Income..k..)

## 0% 25% 50% 75% 100%   
## 15.0 41.5 61.5 78.0 137.0

Standard Deviation

glue("Standard Deviation of Annual Income in $1,000's: {sd(customer\_data$Annual.Income..k..)}")

## Standard Deviation of Annual Income in $1,000's: 26.2647211652712

### Spending Score (1-100)

Min

glue("Min Spending Score: {min(customer\_data$Spending.Score..1.100.)}")

## Min Spending Score: 1

Max

glue("Max Spendinc Score: {max(customer\_data$Spending.Score..1.100.)}")

## Max Spendinc Score: 99

Mean

glue("Mean Spending Score: {mean(customer\_data$Spending.Score..1.100.)}")

## Mean Spending Score: 50.2

Median

glue("Median Spending Score: {median(customer\_data$Spending.Score..1.100.)}")

## Median Spending Score: 50

Quartiles

quantile(customer\_data$Spending.Score..1.100.)

## 0% 25% 50% 75% 100%   
## 1.00 34.75 50.00 73.00 99.00

Standard Deviation

glue("Standard Deviation of Spending Score: {sd(customer\_data$Spending.Score..1.100.)}")

## Standard Deviation of Spending Score: 25.8235216683702

### Observations by Gender

Average age by gender

age\_by\_gender <- aggregate(  
 x = customer\_data$Age,  
 by = list(customer\_data$Gender),  
 FUN = mean  
)  
age\_by\_gender

## Group.1 x  
## 1 Female 38.09821  
## 2 Male 39.80682

Average annual income by gender

income\_by\_gender <- aggregate(  
 x = customer\_data$Annual.Income..k..,  
 by = list(customer\_data$Gender),  
 FUN = mean  
)  
income\_by\_gender

## Group.1 x  
## 1 Female 59.25000  
## 2 Male 62.22727

Average spending score by gender

spending\_by\_gender <- aggregate(  
 x = customer\_data$Spending.Score..1.100.,  
 by = list(customer\_data$Gender),  
 FUN = mean  
)  
spending\_by\_gender

## Group.1 x  
## 1 Female 51.52679  
## 2 Male 48.51136

New DF combining the three above aggregations

summarized\_avgs\_by\_gender <- data.frame(  
 Gender = c("Female", "Male"),  
 Average\_Age = age\_by\_gender$x,  
 Average\_Annual\_Income = income\_by\_gender$x,  
 Avergae\_Spending = spending\_by\_gender$x  
)  
summarized\_avgs\_by\_gender

## Gender Average\_Age Average\_Annual\_Income Avergae\_Spending  
## 1 Female 38.09821 59.25000 51.52679  
## 2 Male 39.80682 62.22727 48.51136

### Summary Stastics

Whole Dataset

describe(customer\_data)

## vars n mean sd median trimmed mad min max range  
## CustomerID 1 200 100.50 57.88 100.5 100.50 74.13 1 200 199  
## Gender\* 2 200 1.44 0.50 1.0 1.43 0.00 1 2 1  
## Age 3 200 38.85 13.97 36.0 37.94 16.31 18 70 52  
## Annual.Income..k.. 4 200 60.56 26.26 61.5 59.64 24.46 15 137 122  
## Spending.Score..1.100. 5 200 50.20 25.82 50.0 50.31 29.65 1 99 98  
## skew kurtosis se  
## CustomerID 0.00 -1.22 4.09  
## Gender\* 0.24 -1.95 0.04  
## Age 0.48 -0.71 0.99  
## Annual.Income..k.. 0.32 -0.15 1.86  
## Spending.Score..1.100. -0.05 -0.86 1.83

By Gender

describe.by(customer\_data, group = customer\_data$Gender)

##   
## Descriptive statistics by group   
## group: Female  
## vars n mean sd median trimmed mad min max range  
## CustomerID 1 112 97.56 58.28 94.5 96.98 73.39 3 197 194  
## Gender\* 2 112 1.00 0.00 1.0 1.00 0.00 1 1 0  
## Age 3 112 38.10 12.64 35.0 37.28 14.83 18 68 50  
## Annual.Income..k.. 4 112 59.25 26.01 60.0 58.29 26.69 16 126 110  
## Spending.Score..1.100. 5 112 51.53 24.11 50.0 51.61 27.43 5 99 94  
## skew kurtosis se  
## CustomerID 0.09 -1.23 5.51  
## Gender\* NaN NaN 0.00  
## Age 0.50 -0.58 1.19  
## Annual.Income..k.. 0.29 -0.48 2.46  
## Spending.Score..1.100. 0.03 -0.81 2.28  
## ------------------------------------------------------------   
## group: Male  
## vars n mean sd median trimmed mad min max range  
## CustomerID 1 88 104.24 57.48 106.5 104.93 68.20 1 200 199  
## Gender\* 2 88 1.00 0.00 1.0 1.00 0.00 1 1 0  
## Age 3 88 39.81 15.51 37.0 39.03 16.31 18 70 52  
## Annual.Income..k.. 4 88 62.23 26.64 62.5 61.39 23.72 15 137 122  
## Spending.Score..1.100. 5 88 48.51 27.90 50.0 48.46 34.10 1 97 96  
## skew kurtosis se  
## CustomerID -0.11 -1.20 6.13  
## Gender\* NaN NaN 0.00  
## Age 0.39 -1.00 1.65  
## Annual.Income..k.. 0.34 0.13 2.84  
## Spending.Score..1.100. -0.06 -1.04 2.97

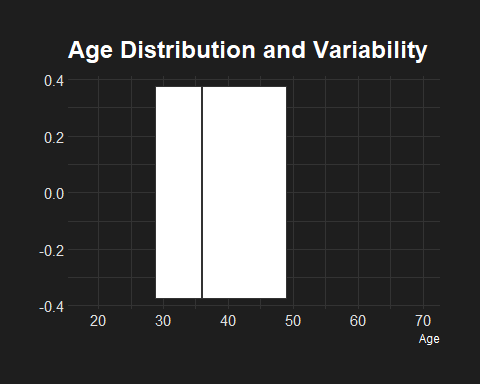
### Data Analysis Summary

In summary, this data set includes 200 observations and consists of data collected by mall customers. 112 of the observations are of females and 88 are of males. The average age of each gender in the data set is about the same at 38.85 years old. On average, the males in this sample earn slightly more with an average of 63,230 dollars per year whereas the average annual income for the females is at 59,250 dollars per year. On average the females spend more with an average spending score of 51.53 out of 100 whereas the males have an average spending score of 48.51.

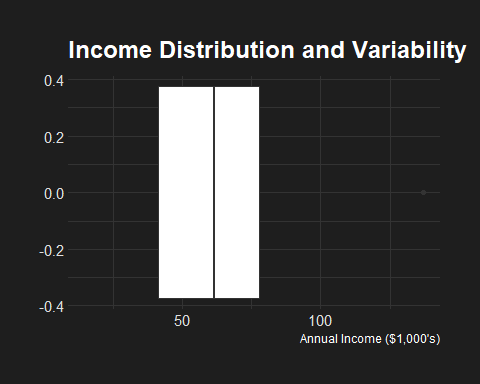
# Plotting

### Boxplots

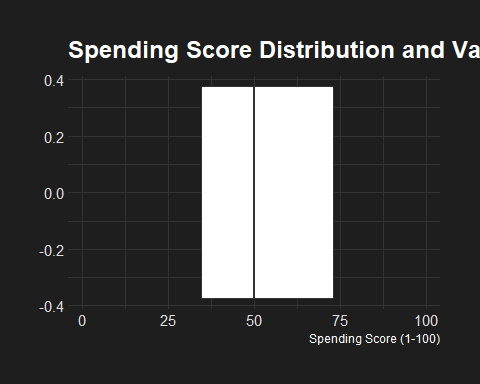
ggplot(data = customer\_data, aes(x = Age)) +  
 geom\_boxplot() +  
 ggtitle("Age Distribution and Variability") +  
 xlab("Age") +  
 theme\_modern\_rc()



ggplot(data = customer\_data, aes(x = Annual.Income..k..)) +  
 geom\_boxplot() +  
 ggtitle("Income Distribution and Variability") +  
 xlab("Annual Income ($1,000's)") +  
 theme\_modern\_rc()

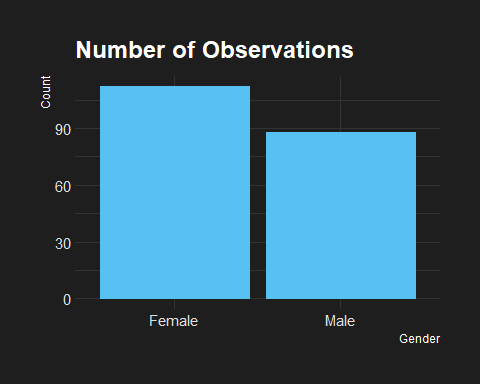


ggplot(data = customer\_data, aes(x = Spending.Score..1.100.)) +  
 geom\_boxplot() +  
 ggtitle("Spending Score Distribution and Variability") +  
 xlab("Spending Score (1-100)") +  
 theme\_modern\_rc()

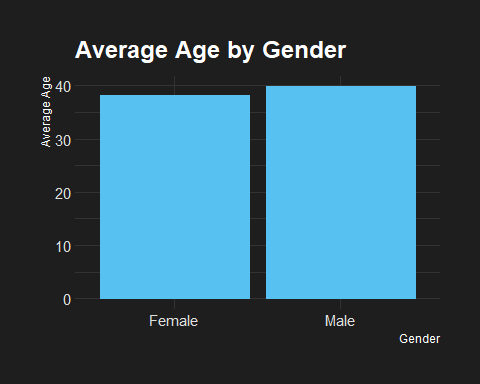


### Histograms (By Gender)

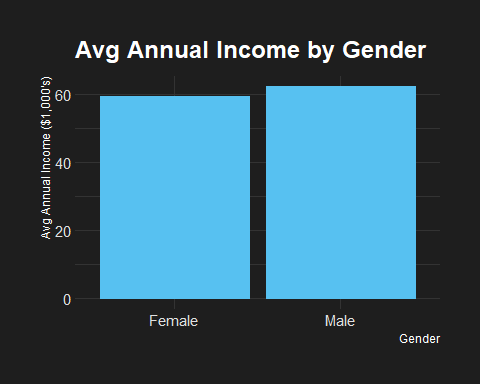
ggplot(data = customer\_data, mapping = aes(x=Gender)) +  
 geom\_bar() +  
 ggtitle ("Number of Observations") +  
 xlab("Gender") +  
 ylab("Count") +  
 theme\_modern\_rc()



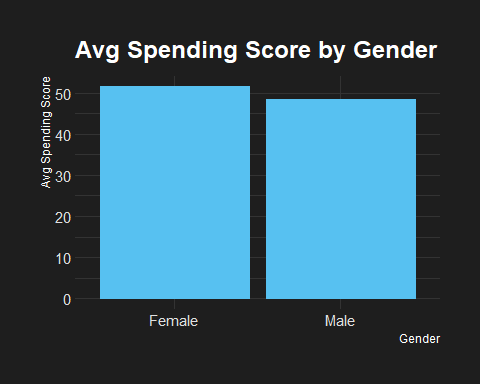
ggplot(data = summarized\_avgs\_by\_gender, mapping = aes(x=Gender, y=Average\_Age)) +  
 geom\_histogram(stat="identity") +  
 ggtitle ("Average Age by Gender") +  
 xlab("Gender") +  
 ylab("Average Age") +  
 theme\_modern\_rc()



ggplot(data = summarized\_avgs\_by\_gender, mapping = aes(x=Gender, y=Average\_Annual\_Income)) +  
 geom\_histogram(stat="identity") +  
 ggtitle ("Avg Annual Income by Gender") +  
 xlab("Gender") +  
 ylab("Avg Annual Income ($1,000's)") +  
 theme\_modern\_rc()



ggplot(data = summarized\_avgs\_by\_gender, mapping = aes(x=Gender, y=Avergae\_Spending)) +  
 geom\_histogram(stat="identity") +  
 ggtitle ("Avg Spending Score by Gender") +  
 xlab("Gender") +  
 ylab("Avg Spending Score") +  
 theme\_modern\_rc()

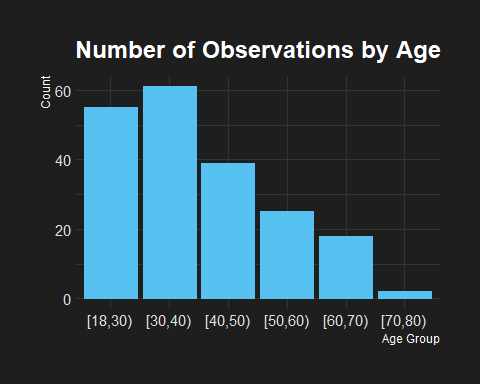


### Histograms (By Age Group)

# ADDING AN AGE GROUP COLUMN TO THE CUSTOMER DATA DF  
age\_cut <- cut(customer\_data$Age, breaks = c(18, 30, 40, 50, 60, 70, 80), right = FALSE)  
customer\_data\_2 <- data.frame(customer\_data, Age\_Group = age\_cut)  
head(customer\_data\_2)

## CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100. Age\_Group  
## 1 1 Male 19 15 39 [18,30)  
## 2 2 Male 21 15 81 [18,30)  
## 3 3 Female 20 16 6 [18,30)  
## 4 4 Female 23 16 77 [18,30)  
## 5 5 Female 31 17 40 [30,40)  
## 6 6 Female 22 17 76 [18,30)

ggplot(customer\_data\_2, aes(x = Age\_Group)) +  
 geom\_bar() +  
 ggtitle ("Number of Observations by Age") +  
 xlab("Age Group") +  
 ylab("Count") +  
 theme\_modern\_rc()



income\_by\_age\_group <- aggregate(  
 x = customer\_data\_2$Annual.Income..k..,  
 by = list(customer\_data\_2$Age\_Group),  
 FUN = mean  
)  
income\_by\_age\_group

## Group.1 x  
## 1 [18,30) 52.54545  
## 2 [30,40) 70.18033  
## 3 [40,50) 63.94872  
## 4 [50,60) 58.68000  
## 5 [60,70) 49.16667  
## 6 [70,80) 47.50000

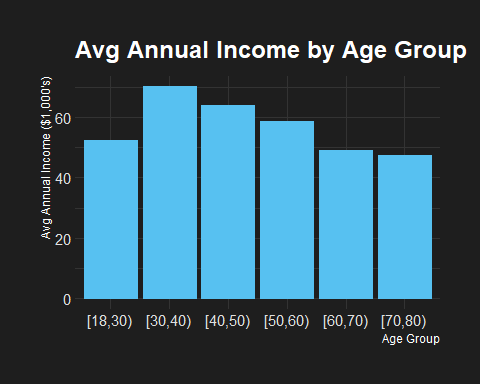
spending\_by\_age\_group <- aggregate(  
 x = customer\_data\_2$Spending.Score..1.100.,  
 by = list(customer\_data\_2$Age\_Group),  
 FUN = mean  
)  
spending\_by\_age\_group

## Group.1 x  
## 1 [18,30) 58.58182  
## 2 [30,40) 61.09836  
## 3 [40,50) 34.94872  
## 4 [50,60) 34.72000  
## 5 [60,70) 41.61111  
## 6 [70,80) 55.50000

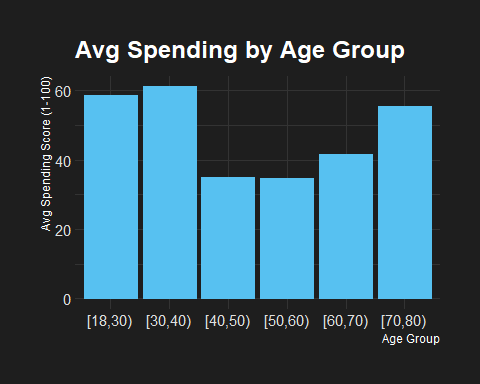
summarized\_avgs\_by\_age\_group <- data.frame(  
 Age\_Group = income\_by\_age\_group$Group.1,  
 Avg\_Income = income\_by\_age\_group$x,  
 Avg\_Spending = spending\_by\_age\_group$x  
)  
summarized\_avgs\_by\_age\_group

## Age\_Group Avg\_Income Avg\_Spending  
## 1 [18,30) 52.54545 58.58182  
## 2 [30,40) 70.18033 61.09836  
## 3 [40,50) 63.94872 34.94872  
## 4 [50,60) 58.68000 34.72000  
## 5 [60,70) 49.16667 41.61111  
## 6 [70,80) 47.50000 55.50000

ggplot(data = summarized\_avgs\_by\_age\_group, mapping = aes(x=Age\_Group, y=Avg\_Income)) +  
 geom\_histogram(stat="identity") +  
 ggtitle ("Avg Annual Income by Age Group") +  
 xlab("Age Group") +  
 ylab("Avg Annual Income ($1,000's)") +  
 theme\_modern\_rc()

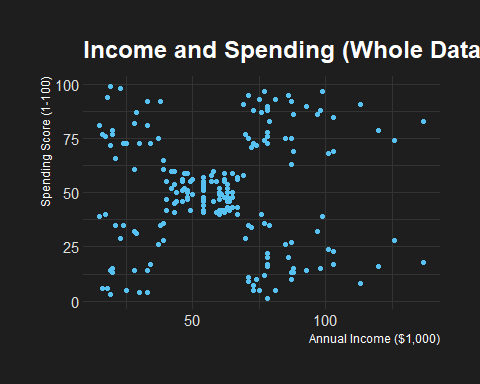


ggplot(data = summarized\_avgs\_by\_age\_group, mapping = aes(x=Age\_Group, y=Avg\_Spending)) +  
 geom\_histogram(stat="identity") +  
 ggtitle ("Avg Spending by Age Group") +  
 xlab("Age Group") +  
 ylab("Avg Spending Score (1-100)") +  
 theme\_modern\_rc()



### Scatterplot

ggplot(data = customer\_data, mapping = aes(x=Annual.Income..k.., y=Spending.Score..1.100.)) +  
 geom\_point() +  
 ggtitle("Income and Spending (Whole Dataset)") +  
 xlab("Annual Income ($1,000)") +  
 ylab("Spending Score (1-100)") +  
 theme\_modern\_rc()



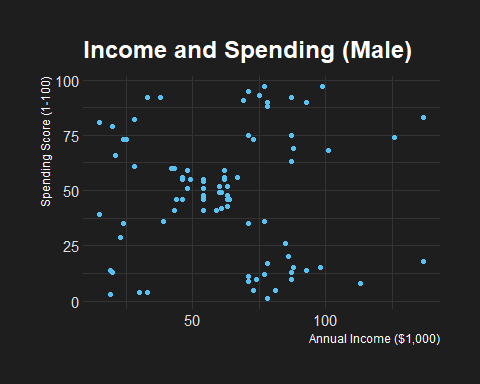
male\_df <- customer\_data\_2[customer\_data\_2$Gender == "Male", ]  
head(male\_df)

## CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100. Age\_Group  
## 1 1 Male 19 15 39 [18,30)  
## 2 2 Male 21 15 81 [18,30)  
## 9 9 Male 64 19 3 [60,70)  
## 11 11 Male 67 19 14 [60,70)  
## 15 15 Male 37 20 13 [30,40)  
## 16 16 Male 22 20 79 [18,30)

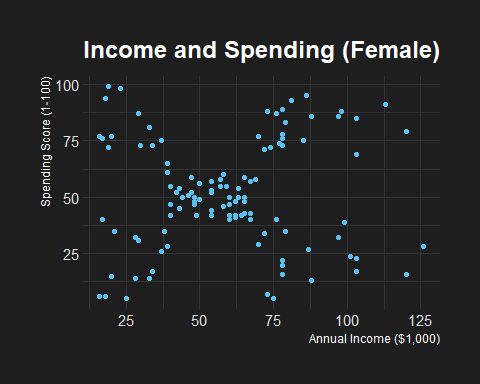
female\_df <- customer\_data\_2[customer\_data\_2$Gender == "Female", ]  
head(female\_df)

## CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100. Age\_Group  
## 3 3 Female 20 16 6 [18,30)  
## 4 4 Female 23 16 77 [18,30)  
## 5 5 Female 31 17 40 [30,40)  
## 6 6 Female 22 17 76 [18,30)  
## 7 7 Female 35 18 6 [30,40)  
## 8 8 Female 23 18 94 [18,30)

ggplot(data = male\_df, mapping = aes(x=Annual.Income..k.., y=Spending.Score..1.100.)) +  
 geom\_point() +  
 ggtitle("Income and Spending (Male)") +  
 xlab("Annual Income ($1,000)") +  
 ylab("Spending Score (1-100)") +  
 theme\_modern\_rc()



ggplot(data = female\_df, mapping = aes(x=Annual.Income..k.., y=Spending.Score..1.100.)) +  
 geom\_point() +  
 ggtitle("Income and Spending (Female)") +  
 xlab("Annual Income ($1,000)") +  
 ylab("Spending Score (1-100)") +  
 theme\_modern\_rc()



### Plot Analysis Summary

From the above plots, we can deduce all of the conclusions mentioned above in addition to the following:

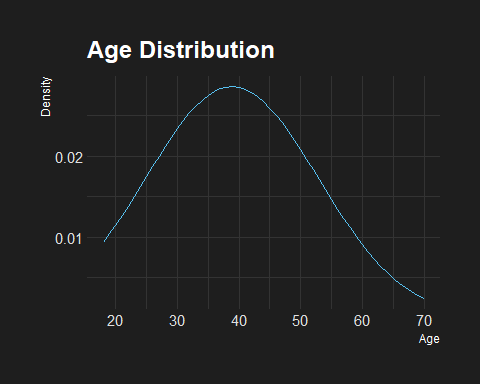
* The age group with the most observations is 30-40 years old with 62 observations
* Well over half of the observations are between ages 18-50
* People between the ages of 30-40 earn the most
* People between the ages 40-50 earn the second most
* People between the ages 50-60 earn the third most
* People between the ages 18-30 and people between the ages 60-70
* People between the ages 18-40 spend the most
* People between the ages 40-60 spend the least
* There is not much of a correlation between income and spending, for both genders

# Probability and Statistical Analysis

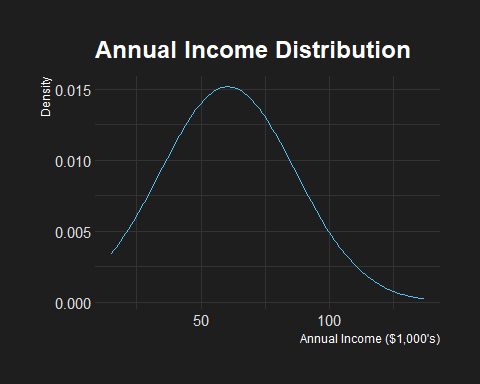
### Distribution

Use the dnorm function to verify whether or not the Age, Annual Income, and Spending Score variables follow the normal distribution

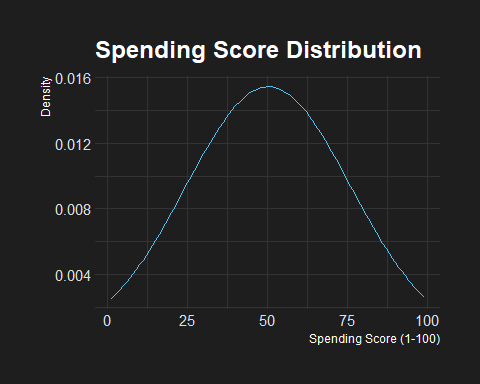
age.range <- min(customer\_data$Age):max(customer\_data$Age)  
age.mean <- mean(customer\_data$Age)  
age.sd <- sd(customer\_data$Age)  
age.dist <- dnorm(age.range, mean = age.mean, sd = age.sd)  
age.df <- data.frame(Age = age.range, Density = age.dist)  
  
ggplot(age.df, aes(x = Age, y = Density)) +  
 geom\_line() +  
 ggtitle("Age Distribution") +  
 theme\_modern\_rc()



income.range <- min(customer\_data$Annual.Income..k..):max(customer\_data$Annual.Income..k..)  
income.mean <- mean(customer\_data$Annual.Income..k..)  
income.sd <- sd(customer\_data$Annual.Income..k..)  
income.dist <- dnorm(income.range, mean = income.mean, sd = income.sd)  
income.df <- data.frame(Income = income.range, Density = income.dist)  
  
ggplot(income.df, aes(x = Income, y = Density)) +  
 geom\_line() +  
 ggtitle("Annual Income Distribution") +  
 xlab("Annual Income ($1,000's)") +  
 theme\_modern\_rc()



spending.range <- min(customer\_data$Spending.Score..1.100.):max(customer\_data$Spending.Score..1.100.)  
spending.mean <- mean(customer\_data$Spending.Score..1.100.)  
spending.sd <- sd(customer\_data$Spending.Score..1.100.)  
spending.dist <- dnorm(spending.range, mean = spending.mean, sd = spending.sd)  
spending.df <- data.frame(Spending = spending.range, Density = spending.dist)  
  
ggplot(spending.df, aes(x = Spending, y = Density)) +  
 geom\_line() +  
 ggtitle("Spending Score Distribution") +  
 xlab("Spending Score (1-100)") +  
 theme\_modern\_rc()



As we can see, all three of these variables (Age, Annual Income, and Spending Score) follow the normal distribution.

### Shapiro-Wilk Normality Test

Shapiro-Wilk normality test to test for normality of each of the variables

* Whole dataset

shapiro.test(customer\_data$Age)

##   
## Shapiro-Wilk normality test  
##   
## data: customer\_data$Age  
## W = 0.95162, p-value = 2.711e-06

shapiro.test(customer\_data$Annual.Income..k..)

##   
## Shapiro-Wilk normality test  
##   
## data: customer\_data$Annual.Income..k..  
## W = 0.97839, p-value = 0.003537

shapiro.test(customer\_data$Spending.Score..1.100.)

##   
## Shapiro-Wilk normality test  
##   
## data: customer\_data$Spending.Score..1.100.  
## W = 0.96946, p-value = 0.0002423

* Male

shapiro.test(male\_df$Age)

##   
## Shapiro-Wilk normality test  
##   
## data: male\_df$Age  
## W = 0.9394, p-value = 0.0004617

shapiro.test(male\_df$Annual.Income..k..)

##   
## Shapiro-Wilk normality test  
##   
## data: male\_df$Annual.Income..k..  
## W = 0.97057, p-value = 0.04253

shapiro.test(male\_df$Spending.Score..1.100.)

##   
## Shapiro-Wilk normality test  
##   
## data: male\_df$Spending.Score..1.100.  
## W = 0.95218, p-value = 0.002627

* Female

shapiro.test(female\_df$Age)

##   
## Shapiro-Wilk normality test  
##   
## data: female\_df$Age  
## W = 0.95635, p-value = 0.001058

shapiro.test(female\_df$Annual.Income..k..)

##   
## Shapiro-Wilk normality test  
##   
## data: female\_df$Annual.Income..k..  
## W = 0.97626, p-value = 0.04328

shapiro.test(female\_df$Spending.Score..1.100.)

##   
## Shapiro-Wilk normality test  
##   
## data: female\_df$Spending.Score..1.100.  
## W = 0.97438, p-value = 0.02977

### Ansari-Bradley Test

Gender-Age

ansari.test(x = customer\_data$Gender, y = customer\_data$Age)

##   
## Ansari-Bradley test  
##   
## data: customer\_data$Gender and customer\_data$Age  
## AB = 20100, p-value = 1  
## alternative hypothesis: true ratio of scales is not equal to 1

Gender-Annual Income

ansari.test(x = customer\_data$Gender, y = customer\_data$Annual.Income..k..)

##   
## Ansari-Bradley test  
##   
## data: customer\_data$Gender and customer\_data$Annual.Income..k..  
## AB = 20100, p-value = 1  
## alternative hypothesis: true ratio of scales is not equal to 1

Gender-Spending Score

ansari.test(x = customer\_data$Gender, y = customer\_data$Spending.Score..1.100.)

##   
## Ansari-Bradley test  
##   
## data: customer\_data$Gender and customer\_data$Spending.Score..1.100.  
## AB = 20100, p-value = 1  
## alternative hypothesis: true ratio of scales is not equal to 1

### Age Probabilities

Probability a person’s age is above 40 in the dataset

glue("{round(pnorm(q = 40, mean = age.mean, sd = age.sd, lower.tail = FALSE) \* 100, digits = 2)}%")

## 46.72%

Probability a person’s age is below 40 in the dataset

glue("{round(pnorm(q = 40, mean = age.mean, sd = age.sd, lower.tail = TRUE) \* 100, digits = 2)}%")

## 53.28%

Probability a person’s age is above 50 in the dataset

glue("{round(pnorm(q = 50, mean = age.mean, sd = age.sd, lower.tail = FALSE) \* 100, digits = 2)}%")

## 21.24%

Probability a person’s age is below 50 in the dataset

glue("{round(pnorm(q = 50, mean = age.mean, sd = age.sd, lower.tail = TRUE) \* 100, digits = 2)}%")

## 78.76%

### corMallData(data,gender) {seperate data into two different DF’s by gender}

corMallData <- function(data = customer\_data, gender) {  
 gDF <- data.frame()  
   
 for (i in 1:200) {  
 if (data[i, 2] == gender) {  
 gData <- cbind(data[i,3],data[i,4], data[i,5])  
 gDF <- rbind(gDF, gData)  
 }  
 }  
   
 return(gDF)  
}  
  
fData <- corMallData(gender = "Female")  
names(fData)[1] <- "Age"  
names(fData)[2] <- "Income"  
names(fData)[3] <- "Spending"  
  
mData <- corMallData(gender = "Male")  
names(mData)[1] <- "Age"  
names(mData)[2] <- "Income"  
names(mData)[3] <- "Spending"  
   
head(fData)

## Age Income Spending  
## 1 20 16 6  
## 2 23 16 77  
## 3 31 17 40  
## 4 22 17 76  
## 5 35 18 6  
## 6 23 18 94

head(mData)

## Age Income Spending  
## 1 19 15 39  
## 2 21 15 81  
## 3 64 19 3  
## 4 67 19 14  
## 5 37 20 13  
## 6 22 20 79

### Probability and Statistical Analysis Summary

* Most of the observations in the data set are of people under the age of 50 (78.76%)
  + We might be able to deduce from this information that most of the people going to malls are under the age of 50
* Gender, Age, Income and Spending Score all have normal distributions

# Overall Summary

It is important to note that this data set only provides a very little amount of information. We can conclude from the data given that the on average, the men earn more than the women and the women spend more than the men. However, we do not have enough information to explain why that might be. Some other variables that might help explain the income difference are education level and career field. One variables that could better help us understand the slight difference in spending among the genders is whether or not the person has children or dependents, and if so, how many. Furthermore, it is important to consider how the smaller amount of people above the age of 50 might be skewing the data. In conclusion, if we want to perform a truly effective analysis, it would be helpful to collect more information to get a better understanding of the situation and be able to explain our findings, rather than simply report them.

The part of this course I enjoyed the most were the labs. Programming is my favorite thing to do in the world and I am very grateful that I have the opportunity to study it and make a career out of it. Prior to this class, I have already had a bit of exposure to R and statistics through self-teaching, so I did not really struggle with any of the concepts in this course. The reason why I took the course was to fulfill a course requirement to graduate. That being said, I would recommend this course to anyone, and I feel that it teaches you a valuable skill set, even for people who aren’t interested in pursuing a career in data analytics or any other career that falls under the umbrella of IT. I’ll see you next semester Dr. Crissman!