Ryan Maciej

Probability & Statistics - West - Homework 13 - EDA

```
library(readr)
library(tidyverse)
library(knitr)
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 80), tidy = TRUE, background = "#CCCCCC")
barterJacks <- read_csv("caseStudy.csv")
origBarterJacks <- read_csv("caseStudy.csv")
```

Part 1 - A

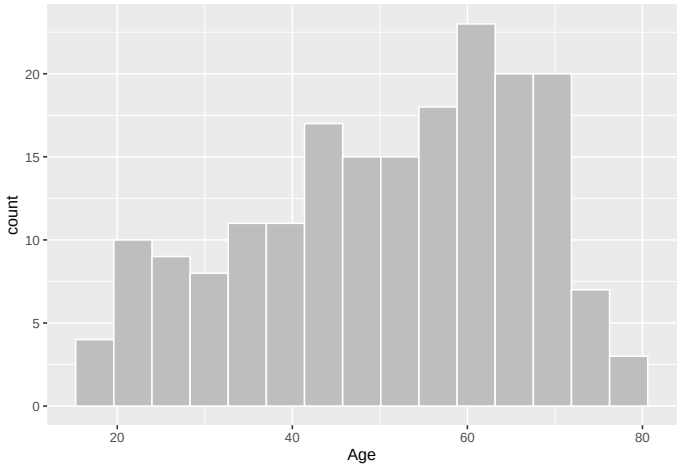
The population that this sample ideally would represent all customers that shop at Barter Jacks in South Bend, although the sampled patrons does not necessarily represent a true random sample due to the fact that certain types of individuals may not would not want to have a frequent shopper card. The implications of this will be discussed in Part 5.

Part 2 - A

Identifiers: Observation, Name
Qualitative Var: Gender, Zip Code, Wine Purchase
Quantitative Var: Age, Purchases, AvgPurchase

Part 2 - B

```
# 2B1
barterJacks %>%
    ggplot(aes(x = Age)) + geom_histogram(bins = 15, color = "white", fill = "gray")
```



```
median(barterJacks$Age)
```

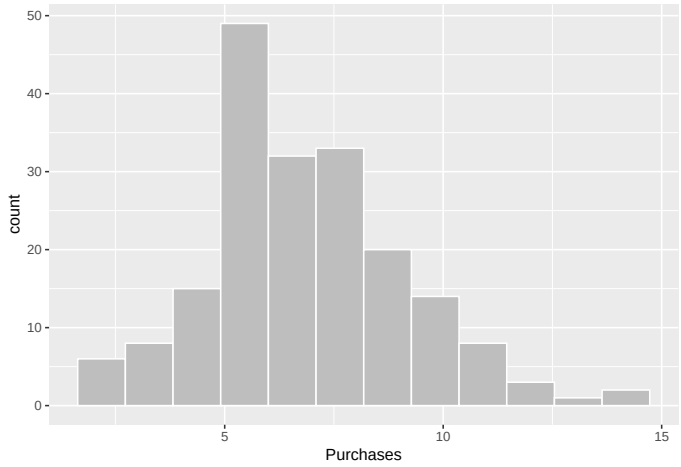
```
## [1] 52
```

```
sd(barterJacks$Age)
```

```
## [1] 15.88024
```

This histogram of age is centered around 52 with a standard deviation of about 16 and is heavily left skewed.

```
barterJacks %>%
    ggplot(aes(x = Purchases)) + geom_histogram(bins = 12, color = "white", fill = "gray")
```




```
median(barterJacks$Purchases)
```

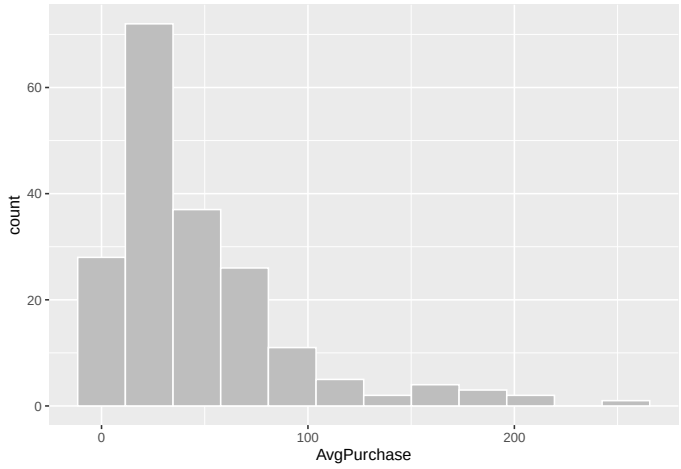
```
## [1] 7
```

```
sd(barterJacks$Purchases)
```

```
## [1] 2.420303
```

This histogram of Purchases is centered around 7 with a standard deviation of about 2.4 and is also slightly left skewed.

```
barterJacks %>%
    ggplot(aes(x = AvgPurchase)) + geom_histogram(bins = 12, color = "white", fill = "gray")
```



```
median(barterJacks$AvgPurchase)
```

```
## [1] 32.35
```

```
sd(barterJacks$AvgPurchase)
```

```
## [1] 44.49005
```

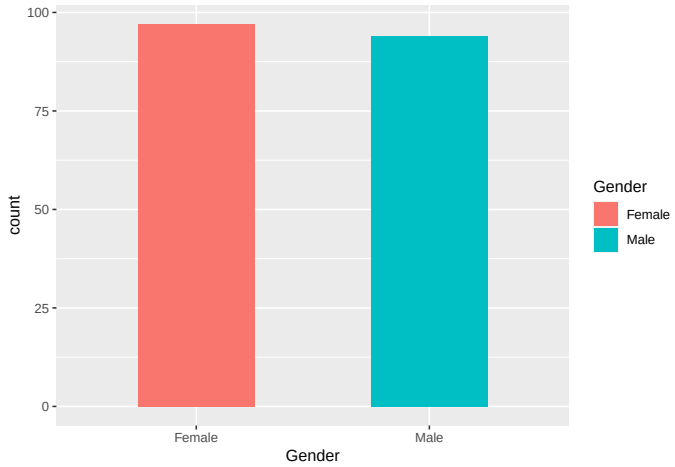
This histogram of Average Purchase is centered around 32 with a standard deviation of about 44 and is heavily right skewed.

Insights: The age of patrons at Barter Jacks with frequent shopper cards tends to be a bit of an older crowd. This could be due to the nature that the older folks have a stronger tendancy to get the frequent shopper card, or that the crowd itself at Barter Jacks is a bit older. Another insight would be that the number of purchases made drops off around 8. With the establishment only opening about 12 weeks ago, it would be interesting to look if there were specific days that were pulling in customers due to events or deals. Lastly, the average purchase tends to be under the 50 Dollar mark but some instances of Average Purchase reach up to the 250 Dollar mark

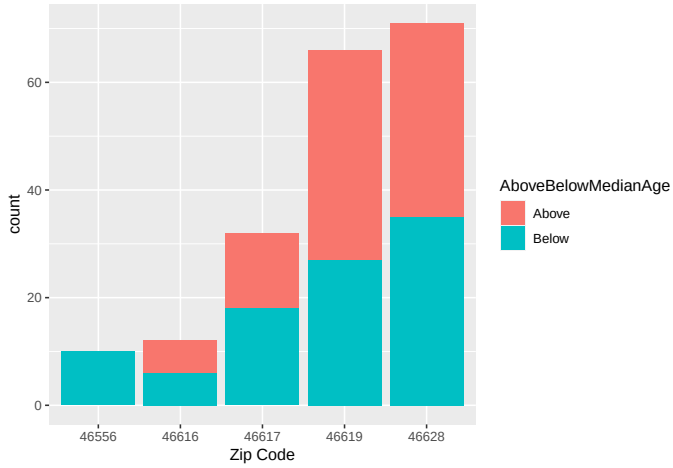
Part 2 - C

```
# Convert Qualitative Values to Character Types
barterJacks$`Zip Code` <- as.character(barterJacks$`Zip Code`)
barterJacks$WinePurchase <- as.character(ifelse(barterJacks$WinePurchase == 1, "Yes",
    "No"))
```

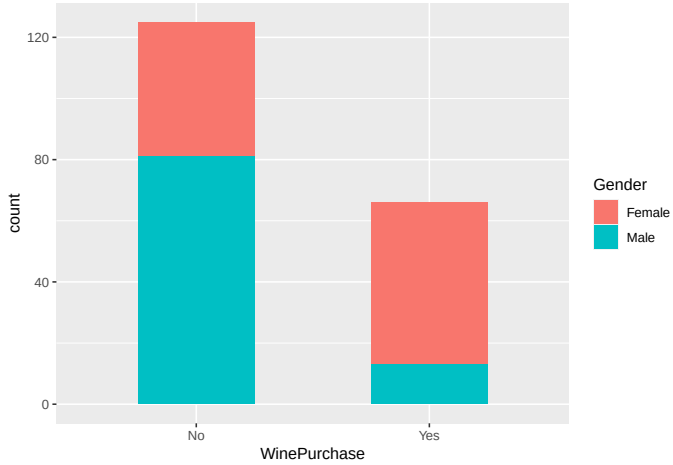
```
barterJacks %>%
    ggplot(aes(x = Gender, fill = Gender)) + geom_bar(width = 0.5)
```




```
medAge <- median(barterJacks$Age)
barterJacks$AboveBelowMedianAge <- ifelse(barterJacks$Age > medAge, "Above", "Below")
barterJacks %>%
    ggplot(aes(x = `Zip Code`, fill = AboveBelowMedianAge)) + geom_bar(position = "stack")
```




```
barterJacks %>%
    ggplot(aes(x = WinePurchase, fill = Gender)) + geom_bar(width = 0.5, position = "stack")
```

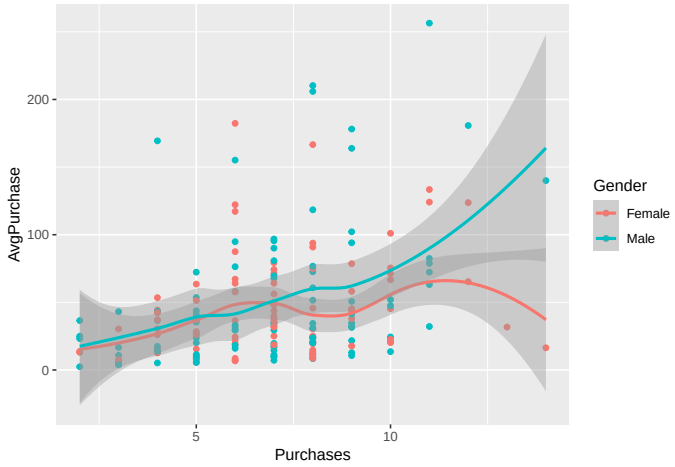


Insights: There appears to be an similar porportion of males as females. Most of the patrons (with the frequent shopper card) tend to be from the zip code 46628. Only about 1/3 of the patrons with cards have made a wine purchase. Most of the wine purchases are made by women. It also looks that all of the patrons (with cards who made purchases) in the zip code of 46556 are below the median age of the sample population.

Part 2 - D

```
barterJacks %>%
    ggplot(aes(x = Purchases, y = AvgPurchase, color = Gender)) + geom_point() +
    geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



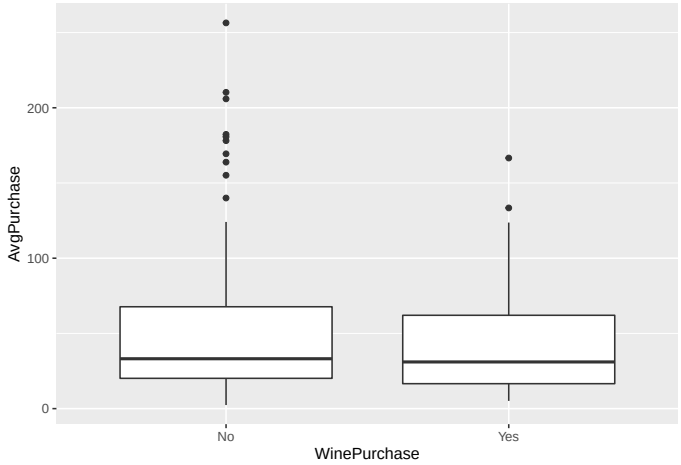

```
cor(x = barterJacks$Purchases, y = barterJacks$AvgPurchase)
```

```
## [1] 0.3300256
```

There looks to be a very slight relationship between number of purchases and average purchase. Breaking it down a bit further between Genders, we see the confidence interval overlapping throughout the whole visualization so we could not confidently make any distinction on Gender and its relation to Average Purchase compared to Purchases.

Part 2 - E

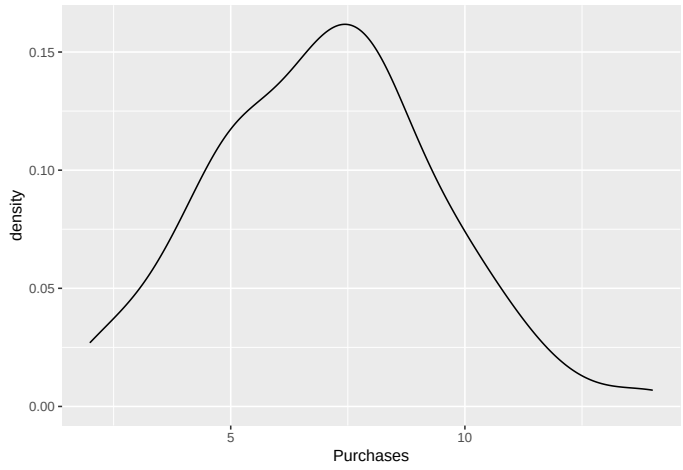
```
barterJacks %>%
    ggplot(aes(WinePurchase, AvgPurchase)) + geom_boxplot()
```



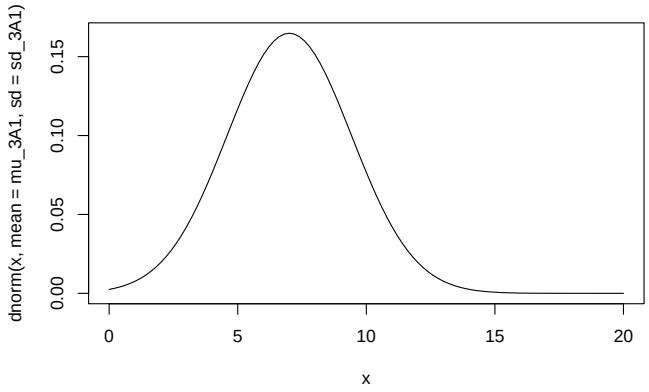
The Box plot shows us that those who purchased wine have very similar avg purchases that those who have not purchased wine. From the bar chart above we know that more people have not bought wine than have bought wine which might explain a bit why they have more outliers.

Part 3 - A

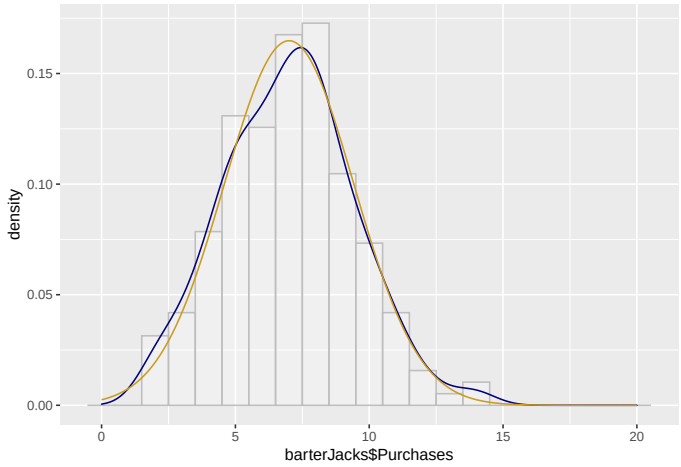
```
barterJacks %>%
    ggplot() + geom_density(aes(Purchases))
```




```
mu_3A1 <- mean(barterJacks$Purchases)
var_3A1 <- var(barterJacks$Purchases)
sd_3A1 <- sd(barterJacks$Purchases)

curve(dnorm(x, mean = mu_3A1, sd = sd_3A1), from = 0, to = 20)
```




```
x_3A1 <- seq(0, 20, length = 300)
dn_3A1 <- dnorm(x_3A1, mean = mu_3A1, sd = sd_3A1)

ggplot() + geom_histogram(aes(x = barterJacks$Purchases, y = ..density..),
    binwidth = 1, color = "gray", fill = "white", alpha = 0.3) +
    geom_density(aes(barterJacks$Purchases), color = "navy") +
    geom_line(aes(x_3A1, dn_3A1), color = "goldenrod3")
```

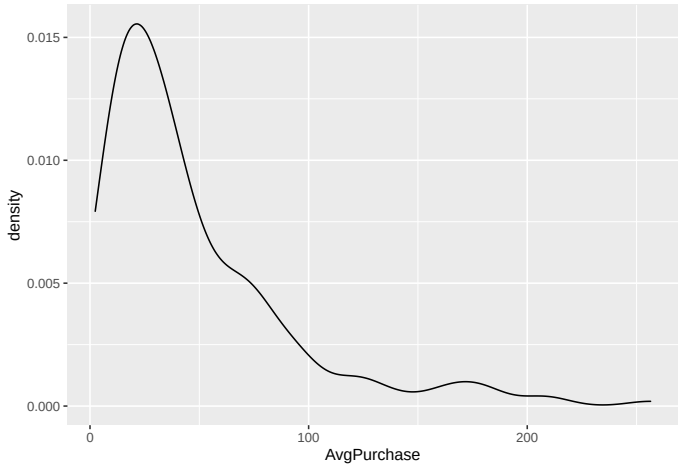


Purchases appears to be a normal distribution with mean of 7 and standard deviation of 2.42

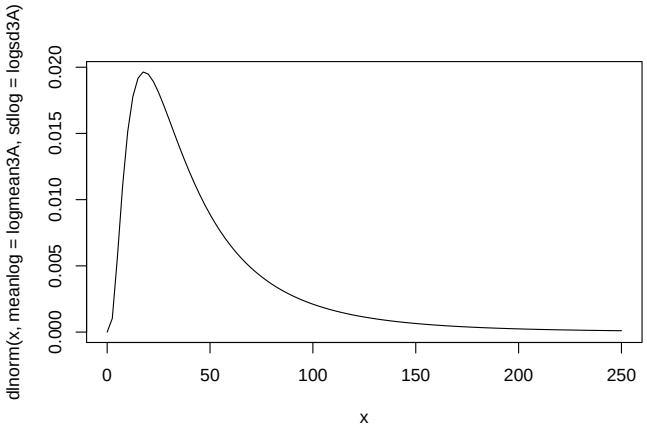
Part 3 - A (Extra Credit?!?!)

Will be using a Lognormal Distribution with the Avg. Purchase Density plot.

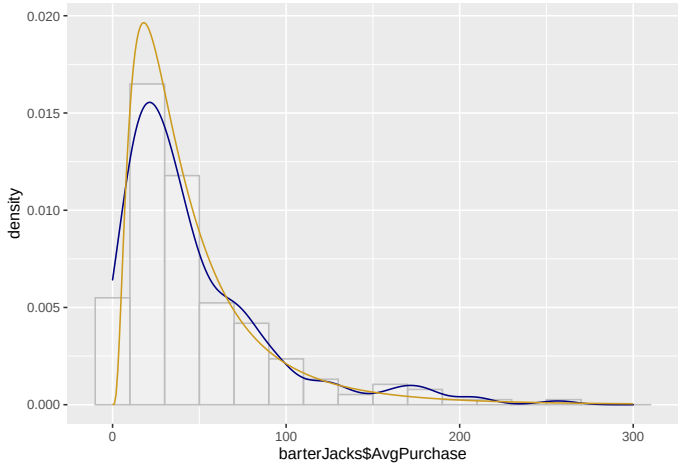
```
barterJacks %>%
    ggplot() + geom_density(aes(AvgPurchase))
```




```
mu_3A <- mean(barterJacks$AvgPurchase)
var_3A <- var(barterJacks$AvgPurchase)
sd_3A <- sd(barterJacks$AvgPurchase)

logmean3A <- log(mu_3A^2/sqrt(sd_3A^2 + mu_3A^2))
logsd3A <- sqrt(log(1 + (sd_3A/mu_3A)))

curve(dlnorm(x, meanlog = logmean3A, sdlog = logsd3A), from = 0, to = 250)
```




```
x_3A <- seq(0, 300, length = 300)
dn_3A <- dlnorm(x_3A, meanlog = logmean3A, sdlog = logsd3A)

ggplot() + geom_histogram(aes(x = barterJacks$AvgPurchase, y = ..density..), binwidth = 20,
    color = "gray", fill = "white", alpha = 0.3) + geom_density(aes(barterJacks$AvgPurchase),
    color = "navy") + geom_line(aes(x_3A, dn_3A), color = "goldenrod3")
```



Part 3 - B

```
barterJacks$AgeAdj <- barterJacks$Age/max(barterJacks$Age)
A <- min(barterJacks$AgeAdj)
B <- max(barterJacks$AgeAdj)
A + (B - A)
```

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

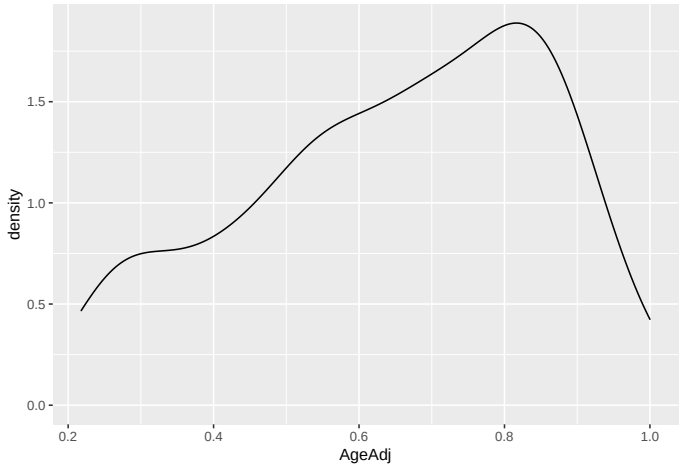
```
range3B <- (B - A)^2
mu_AgeAdj <- mean(barterJacks$AgeAdj)
var_AgeAdj <- var(barterJacks$AgeAdj)
mu_AgeAdj
```

```
## [1] 0.6490804
```

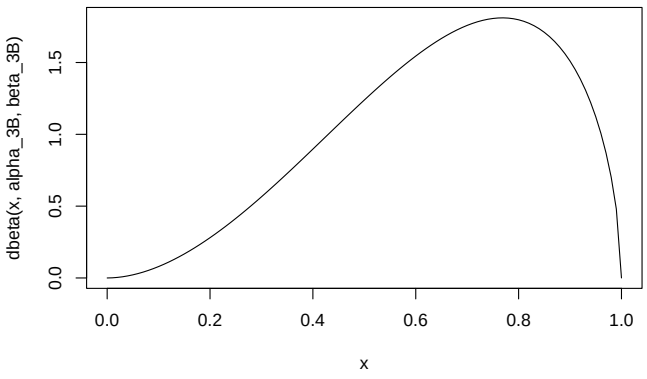
```
var_AgeAdj
```

```
## [1] 0.04145006
```

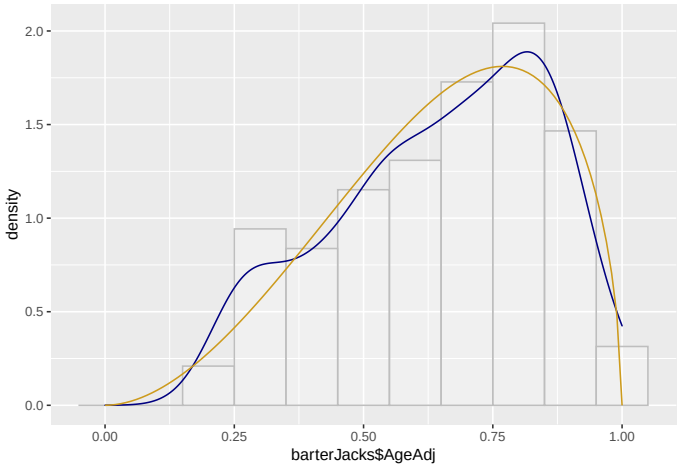
```
barterJacks %>%
    ggplot(aes(x = AgeAdj)) + geom_density()
```




```
# See Mathematica Code after R Code for alpha and beta calculation (Roughly
# Page 20)

alpha_3B <- 2.91773
beta_3B <- 1.5774

x <- seq(0, 1, length = 100)
db <- dbeta(x, alpha_3B, beta_3B)

curve(dbeta(x, alpha_3B, beta_3B))
```




```
ggplot() + geom_histogram(aes(x = barterJacks$AgeAdj, y = ..density..), binwidth = 0.1,
    color = "gray", fill = "white", alpha = 0.3) + geom_density(aes(x = barterJacks$AgeAdj),
    color = "navy") + geom_line(aes(x, db), color = "goldenrod3")
```


density
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
barterJacks$AgeAdj

Part 4 - A

```
# Create new column of Total Purchase
barterJacks$TotalPurchase <- barterJacks$Purchases * barterJacks$AvgPurchase

zip46628 <- filter(barterJacks, barterJacks$`Zip Code` == "46628")
n <- length(zip46628$Observation)
x_bar <- mean(zip46628$TotalPurchase)
se <- sd(zip46628$TotalPurchase)

testStat_4A <- (x_bar - 350)/(se/sqrt(n))
p_4A <- 1 - pt(testStat_4A, n - 1)
p_4A
```

```
## [1] 0.2600022
```

```
t.test(zip46628$TotalPurchase, mu = 350, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  zip46628$TotalPurchase
## t = 0.6466, df = 70, p-value = 0.26
## alternative hypothesis: true mean is greater than 350
## 95 percent confidence interval:
##  293.9528      Inf
## sample estimates:
## mean of x
##  385.5187
```

Testing to see if the true population mean of Total Purchase for patrons in 46628 is greater than $350 This is a single sample looking at the population mean.

Conditions of Inference are met since n is large.

Null: The population mean is less than or equal to the 350. Alternate: The population mean is greater than 350.

Test Statistic: .6466 p-value: .26

$p > .05$ therefore we fail to reject the Null.

There is not sufficient evidence to suggest that the true population mean of patrons with a zipcode of 46628 is above $350

Part 4 - B

```
barterJacks4B1 <- filter(barterJacks, barterJacks$Purchases >= 10 & (barterJacks$`Zip Code` ==
    "46617" | barterJacks$`Zip Code` == "46556"))
barterJacks4B2 <- filter(barterJacks, barterJacks$`Zip Code` == "46617" | barterJacks$`Zip Code` ==
    "46556")
table(barterJacks4B1$`Zip Code`)
```

```
##
## 46556 46617
##     2     4
```

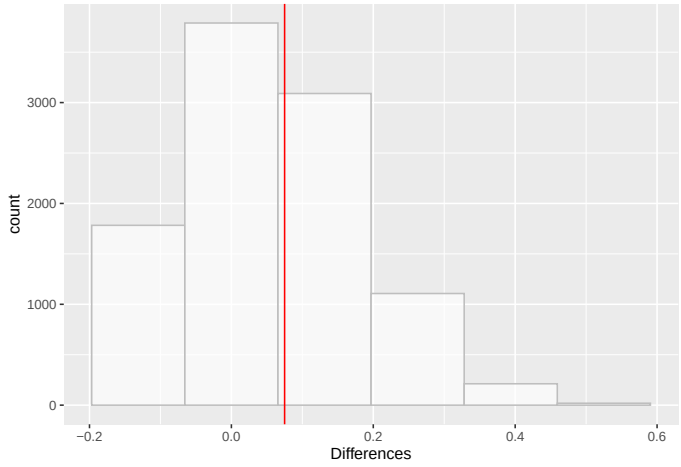
```r
table(barterJacks4B2$`Zip Code`)
```

```
##
## 46556 46617
##    10    32
```

```r
Ten46556 <- c(rep(1, 2), rep(0, 8))
Ten46617 <- c(rep(1, 4), rep(0, 28))

bothZip <- c(Ten46556, Ten46617)
group <- c(rep("46556", 10), rep("46617", 32))

zipData <- data.frame(cbind(bothZip, group))

origDiff_4B <- mean(Ten46556) - mean(Ten46617)
origDiff_4B
```

```
## [1] 0.075
```

```r
iters <- 10000
diff_4B <- c(rep(NA, iters))

temp_4B <- group
set.seed(1987)
for (i in 1:iters) {
    temp_4B <- sample(temp_4B, length(group), replace = F)
    diff_4B[i] <- sum(temp_4B == "46556" & bothZip == 1)/length(Ten46556) - sum(temp_4B ==
        "46617" & bothZip == 1)/length(Ten46617)
}

diff_4B <- data.frame(diff_4B)
colnames(diff_4B) <- "Differences"

diff_4B %>%
    ggplot(aes(x = Differences)) + geom_histogram(color = "gray", fill = "white",
    alpha = 0.7, bins = 6) + geom_vline(xintercept = origDiff_4B, color = "red")
```




```r
p_4B <- sum(diff_4B >= origDiff_4B)/iters
p_4B
```

```
## [1] 0.4428
```

Testing to see if there is a difference in true population proportions based on the sample.

Conditions of inference were not met, because np, n(1-p) were not all >= 10 for both samples. A simulation was conducted to determine an appropriate p value which came out to be .4428

Our Null would be that patrons from 46556 have the same or less proportion of individuals who have at least 10 purchases than 46617 Out Alternate would be that patrons from 46556 have a higher proportion of individuals who have at least 10 purchases than 46617

Because we have a large p-value, we fail to reject the full and there is no evidence to believe patrons from 46556 have a higher proportion of individuals who have at least 10 purchases.

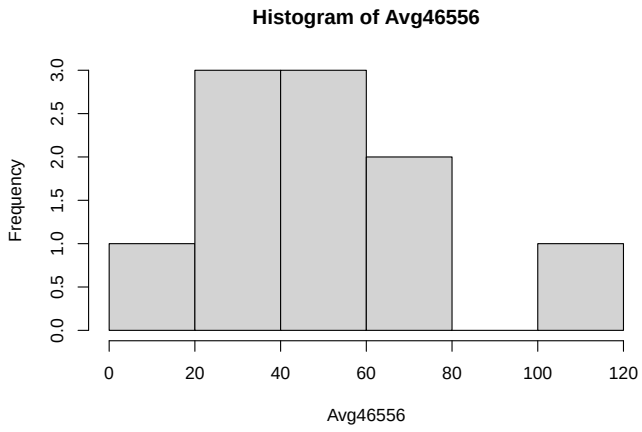
Part 4 - C

Finding if there is a difference between the population means of two zip codes, 46617 & 46556.

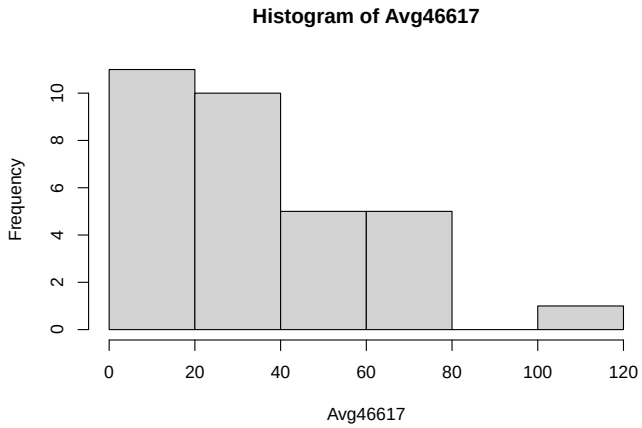
```
barterJacks4C <- filter(barterJacks, barterJacks$`Zip Code` == "46617" | barterJacks$`Zip Code` ==
    "46556")
barterJacks4C_summary <- barterJacks4C %>%
    group_by(`Zip Code`) %>%
    summarise(COUNT = n(), MEAN = mean(AvgPurchase), SD = sd(AvgPurchase), VAR = var(AvgPurchase))
barterJacks4C_summary
```

```
## # A tibble: 2 x 5
##   `Zip Code` COUNT  MEAN    SD   VAR
##   <chr>      <int> <dbl> <dbl> <dbl>
## 1 46556         10  50.2  26.5  700.
## 2 46617         32  36.2  23.9  572.
```

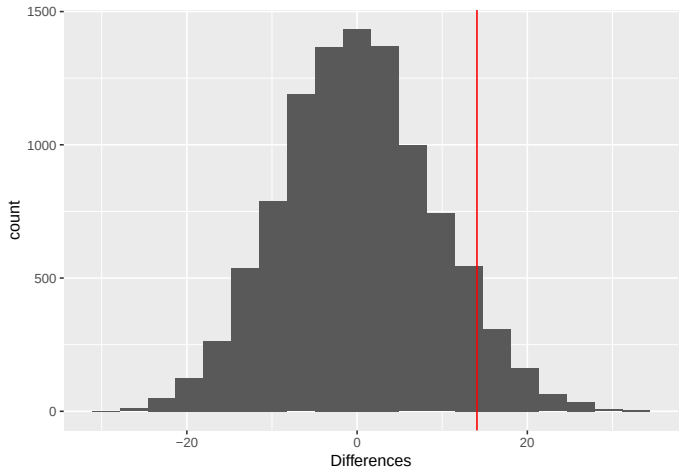
```
Avg46617 <- barterJacks %>%
    filter(`Zip Code` == "46617") %>%
    select(AvgPurchase) %>%
    .$AvgPurchase
Avg46556 <- barterJacks %>%
    filter(`Zip Code` == "46556") %>%
    select(AvgPurchase) %>%
    .$AvgPurchase
hist(Avg46556)
```



```
hist(Avg46617)
```




```
Zip46617 <- c(rep("46617", length(Avg46617)))
Zip46556 <- c(rep("46556", length(Avg46556)))

Avg4C <- c(Avg46617, Avg46556)
Zip4C <- c(Zip46617, Zip46556)

origDiff <- (barterJacks4C_summary %>%
    filter(`Zip Code` == "46556") %>%
    select(MEAN)) - (barterJacks4C_summary %>%
    filter(`Zip Code` == "46617")) %>%
    select(MEAN)
origDiff2 <- unlist(origDiff$MEAN)

temp <- Zip4C
iters <- 10000
diffs <- c(rep(NA, iters))

set.seed(1912)
for (i in 1:iters) {
    temp <- sample(Zip4C, length(Zip4C), replace = F)
    diffs[i] <- mean(Avg4C[which(temp == "46556")]) - mean(Avg4C[which(temp == "46617")])
}

diffs <- data.frame(diffs)
colnames(diffs) <- "Differences"

ggplot(diffs, aes(x = Differences)) + geom_histogram(bins = 20) + geom_vline(xintercept = origDiff2,
    color = "red")
```




```
2 * sum(diffs > origDiff2)/iters
```

```
## [1] 0.1342
```

```
var(Avg46556)/var(Avg46617)
```

```
## [1] 1.222427
```

```
pooled4C <- t.test(Avg46617, Avg46556, var.equal = T)
pooled4C$p.value
```

```
## [1] 0.1207476
```

```
unpooled4C <- t.test(Avg46617, Avg46556, var.equal = F)
unpooled4C$p.value
```

```
## [1] 0.1553576
```

Quick Look at Total Purchase

```
Tot46617 <- barterJacks %>%
    filter(`Zip Code` == "46617") %>%
    select(TotalPurchase) %>%
    .$TotalPurchase
Tot46556 <- barterJacks %>%
    filter(`Zip Code` == "46556") %>%
    select(TotalPurchase) %>%
    .$TotalPurchase
var(Tot46556)/var(Tot46617)
```

```
## [1] 1.06949
```

```
pooled4C2 <- t.test(Tot46617, Tot46556, var.equal = T)
pooled4C2$p.value
```

```
## [1] 0.1858578
```

We are looking to see if there is a difference in the average amount spent between two zip codes 46556 and 46617

This would be a two sample, two tailed test of population means.

Because our n1 and n2 aren't both large, I decided to do a simulation and compare it to the theoretical values that that t.test function provides, both with var.eqaul being T & F.

From the simulation, we were able to obtain a p-value of .1342 Because the variances are close enough to be considered the same, the val.equal = T option would be best, yeilding a p-value of .1207 I also checked the p-value if we did not want to conclude the variance were the same and got a p-value of .1554

Because all 3 of these results are $> .05$ we again fail to reject our Null (The Null Being there is No Difference) and can conclude there is not sufficent evidence to believe that there is any differece in the population means of Average Purchase.

A reasonable followup question would be the same question, but the total amount purchased. These variances are even closer than the Avg Purchase samples and yield a p-value of .1859

Therefor there is not sufficient evidence to believe that there is any difference in either Avg Purchase or Total Spend between the two zip codes.

Part 4 - D - I

```
contTable <- barterJacks %>%
    group_by(Gender, WinePurchase) %>%
    summarise(Count = n()) %>%
    pivot_wider(names_from = WinePurchase, values_from = Count)
```

```
## `summarise()` has grouped output by 'Gender'. You can override using the `.groups` argument.
```

```
contTable
```

```
## # A tibble: 2 x 3
## # Groups:   Gender [2]
##   Gender    No   Yes
##   <chr>  <int> <int>
## 1 Female    44    53
## 2 Male      81    13
```

```
n11 <- contTable[[1, 2]]
n12 <- contTable[[1, 3]]
n21 <- contTable[[2, 2]]
n22 <- contTable[[2, 3]]

# Odds Ratios: Men to Women
oddR_MW <- (n11 * n22)/(n12 * n21)

# Women to Men
oddR_WM <- 1/((n11 * n22)/(n12 * n21))
```

The odds of making a purchase which includes wine by a lady are about 7.5 times that of the gentlemen.

Part 4 - D - II

```
CI_4D <- c(0, 0)
CI_4D[1] <- log(oddR_WM) - 2.576 * sqrt(1/n11 + 1/n12 + 1/n21 + 1/n22)
CI_4D[2] <- log(oddR_WM) + 2.576 * sqrt(1/n11 + 1/n12 + 1/n21 + 1/n22)
CI_4D
```

```
## [1] 1.083731 2.947473
```

We are 99% confident that the log odds for a purchase including wine for women vs men are between 1.08 and 2.95. Since zero does not fall in this interval, there is evidence of a relationship between genders and the liklihood this purchase wine.

Part 4 - D - III

```
x_4D3 <- data.matrix(contTable[c("No", "Yes")])
rownames(x_4D3) <- c("F", "M")
x_4D3
```

```
##   No Yes
## F 44  53
## M 81  13
```

```
test_4D3 <- chisq.test(x_4D3, correct = F)
test_4D3
```

```
##
##  Pearson's Chi-squared test
##
## data:  x_4D3
## X-squared = 35.156, df = 1, p-value = 3.043e-09
```

With a very low p-value, we can reject the null and conclude Wine Purchase and Gender are not Independent.

In[7]:=
```
Solve[Divide[α, α + β] == .6490804 &&
  Divide[(α * β), Power[α + β, 2] * (α + β + 1)] == .04145006, {α, β}]
```

Out[7]=
$$\{\{\alpha \to 2.91773,\ \beta \to 1.57744\}\}$$

Part 5

After doing an initial analysis of the Barter Jacks sales data, there are a few things we should address before diving too deep into how some of the given data might help us improve sales and bring more customers into the store. These include how the data collection process might be improved to give us more accuracy and insight into the spending habits of the patrons as well as addressing potential biases that may affect the sample resulting in a less than perfect representation of the population of patrons as whole.

The main item of concern that would introduce bias to the sample population is that we are only sampling patrons who have joined the FSC (Frequent Shopper Card) program. There could potentially be a subset of people who have similar characteristics or spending habits that opt out of the program whose data we aren't including in this analysis. Examples of this could be that men are more likely to decline signing up for the FSC than women. Similarly, we don't have information as to whether specific customers are using their FSCs each time they make a purchase. Allowing patrons to use a phone number, or other identifier, in the event they forgot their card may ensure more consistency in its use.

Another aspect of the data that could be improved upon to help with the analysis would be increasing the granularity of the data. If the FSC program could create observations based on each purchase, as opposed to each individual, we could easily look deeper into the data to extract useful information. This setup would allow us to obtain the information we are already collecting with aggregate functions but add new options in ways we can look at the data.

With all the being said we were able to look into the data provided and obtain some valuable information that should help improve sales and bring more people into the store.

One of the items mentioned about Barter Jacks was that is known for their inexpensive wines. We looked specifically at Gender and how it related to those making the most wine purchases and are able to conclude that there is a relationship between genders and who is more likely to purchase wine. The bar chart on page 4 helps visualize that difference and our statistical analysis of the two sample populations (men vs women) on page 19 (Part 4 – A) reiterates what we see in the chart at a glance concluding that there is evidence of a relationship, specifically that women are much more likely to purchase wine than men.

A few of the tests we conducted didn't provide us with much as much actionable items specifically some of the spending habits related to zip code. We could not provide evidence to support that the total average spent from patrons from 46628 was over $350 even though our

sample mean was closer to $385. We could not provide evidence to believe that there was a higher proportion of patrons with 10 purchases or from the zip code 46556 then 46617. Nor were we able to provide evidence that there was a difference in average amount spent between the same two zip codes.

Given more time, and the continued collection of data from the FSC program, we would be able to revisit these various analysis. A larger set of data will give us more accurate results given that the data collection process remains the same or improves in ways mentioned above. One of the major hurdles in data analysis is inconsistent data collected over time. With the suggestions on how to better improve your data collection process, we hope we can continue to work with you to create a methodology that will help provide the most insight overtime. We are confident that the analysis provided will give you actionable insight to help both improve sales and bring customers into your establishment.