

Probability and Statistics HW13


Name = Srikanth Vishnuvajhala

Section = East

Problem Introduction

Data

Barter Jacks recently opened a location in South Bend, Indiana, close to Notre Dame's campus. They ask patrons to sign up for their frequent shopper card to learn more about their customer's shopping patterns. Since opening 3 months ago, they have had 191 patrons sign up for their frequent shopper card. They have recruited you to make sense of some of the data that they have collected so far.

The dataset can be found [here](#) , and contains the following variables:

- Gender – Whether the patron identifies as male or female
- Age – The patron's current age, calculated by the birthdate they provided (not in the data set)
- Zip Code – The zip code for the address that the patron provided
- Purchases – The number of times the frequent shopper card has been scanned for a purchase
- AvgPurchase – The average amount that the patron has spent on purchases while using their frequent shopper card.
- WinePurchase – An indicator variable as to whether the patron has purchased a bottle of wine in any of the transactions where their frequent shopper card has been scanned.

Define the Problem

When the manager gave you the data, she told you “My hope for this project is to gain some insight into my customers, improve sales and bring more people into the store.”

The following questions are intended to get you started on exploring the data and provide you with information that you could share with the manager. These are the *minimum* elements for this case study. If you feel you need to do more exploration and testing in order to adequately meet the needs of the manager, then you should.

What to Complete

1. Defining the Population of Interest

A. What is the population that this sample represents? That is, to whom can we make generalizations using this data?

Answer to Question:1 = The population that this sample represents are people who sign up for the frequent shopper card is South Bend, Indiana. We have taken a sample of 191 people that have frequent shopper card. From this sample of 191 people, we must make a generalization towards the population of South Bend, Indiana who sign up for the frequent shopper card.

2. Exploratory Data Analysis

A. For each of the variables: Gender, Age, Zip Code, Purchases, Average Purchase and Wine Purchases, determine if the variable is quantitative or qualitative.

Answer to Question:2-A =

Variable	Qualitative/ Quantitative
Gender	Qualitative
Age	Quantitative
Zip Code	Qualitative
Purchases	Quantitative
Average Purchase	Quantitative
Wine Purchases	Qualitative

Solutions for the other questions in the next few pages.

2. Exploratory Data Analysis

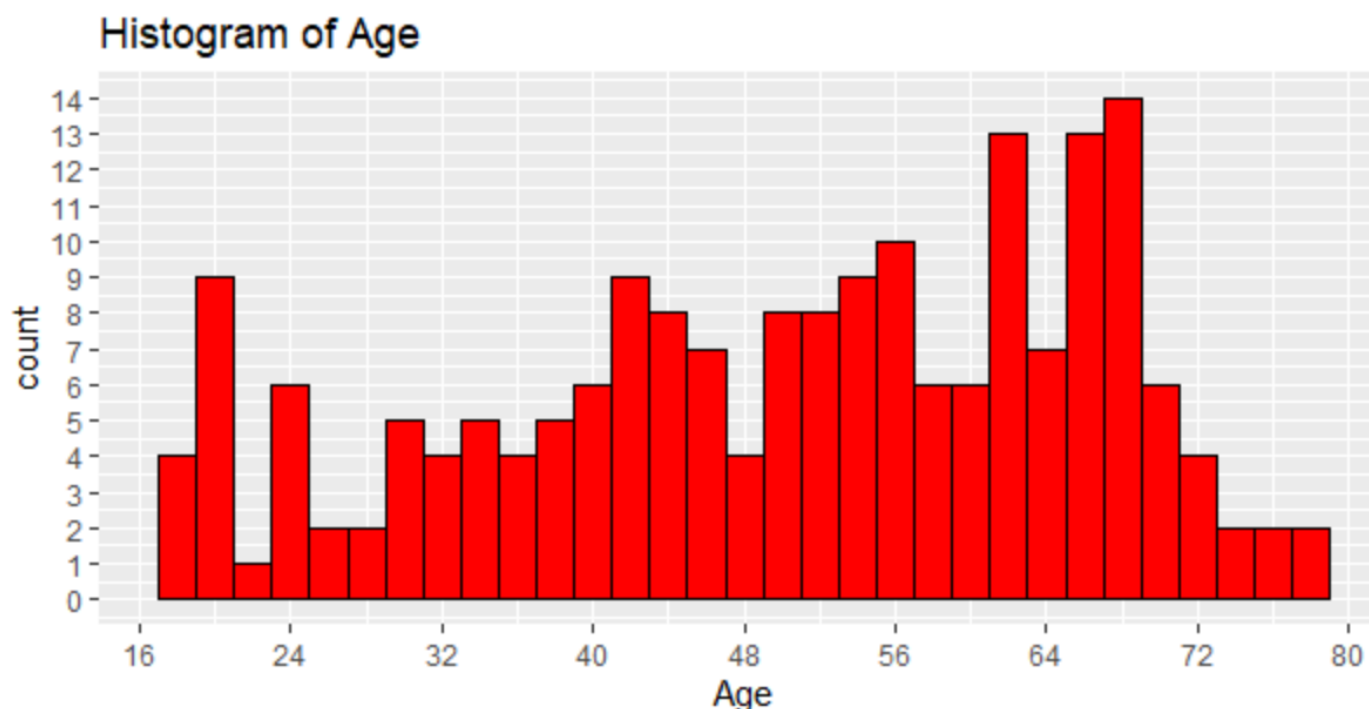
B. Create a histogram and describe the distribution (shape, center and spread) for each of your quantitative variables. Provide at least 3 insights that these graphs have provided you.

Answer to Question:2-B =

Quantitative Variable - 1 = Age

```
ggplot(  
  data = data_purchases,  
  aes(x = Age)  
) +  
  geom_histogram(color = "black", fill = "red", binwidth = 2) +  
  ggtitle("Histogram of Age") +  
  scale_x_continuous(  
    name = "Age",  
    breaks = seq(16,110,8),  
    labels = seq(16,110,8)) +  
  scale_y_continuous(  
    name = "count",  
    breaks = seq(0,20,1),  
    labels = seq(0,20,1))
```

Histogram =



```
# Shape = Left Skewed  
  
# Center =  
mean(data_purchases$Age)  
# Output = 50.62827 = 51 (approx)  
median(data_purchases$Age)  
# Output = 52  
sd(data_purchases$Age)
```

```
# Output = 15.88024 = 16 (approx)

# Spread of data is
(51-16)
# Output = 35
(51+16)
# Output = 67
# The spread of data is between (35,67) Age Group
```

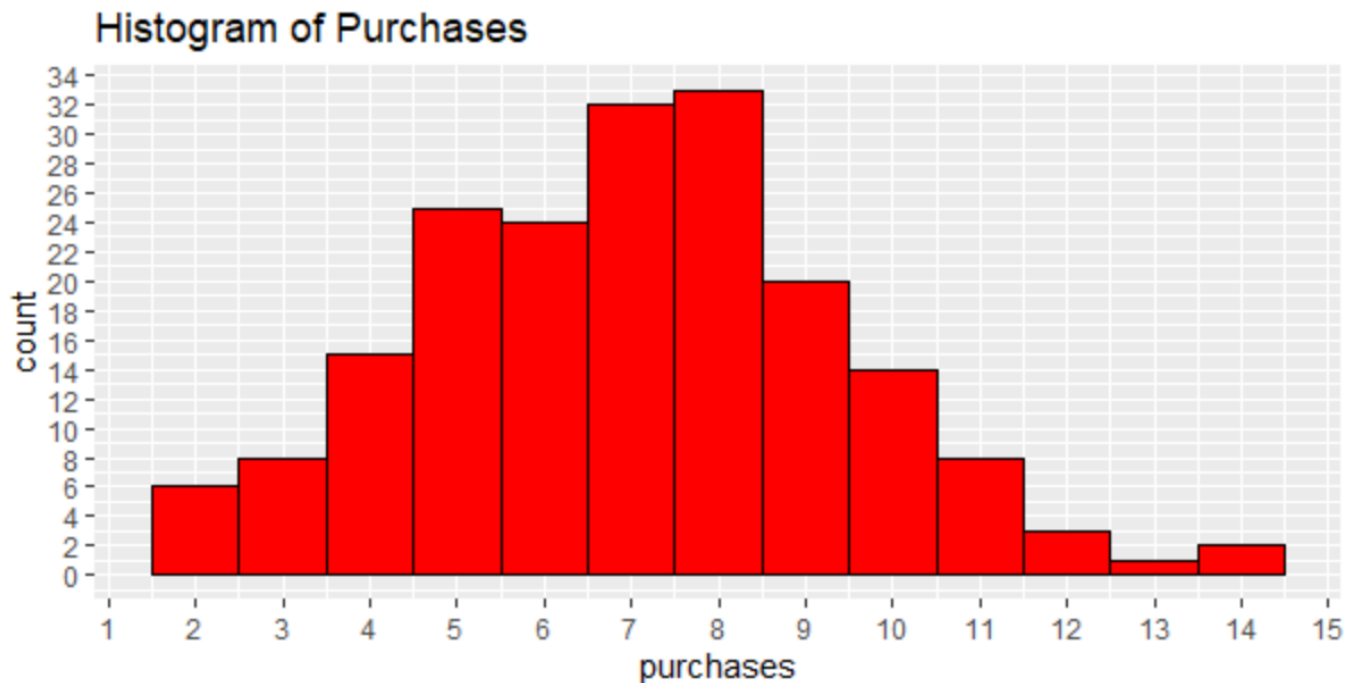
Insights from the Histogram of Age =

1. People from age 17 to age 78 have bought at the store using the Shoppers card.
 2. You see more peaks after age 60 which means people from age 60 to 68 have bought at the store more often using the shoppers card.
 3. The most loyal customers using the shoppers card have been between ages 35 to 67.
 4. The mean is 50.62827, median is 52 and standard deviation is 15.88024.
-

Quantitative Variable - 2 = Purchases

```
library(ggplot2)
ggplot(
  data = data_purchases,
  aes(x = Purchases)
) +
  geom_histogram(color = "black", fill = "red", binwidth = 1) +
  ggtitle("Histogram of Purchases") +
  scale_x_continuous(
    name = "purchases",
    breaks = seq(1,15,1),
    labels = seq(1,15,1)) +
  scale_y_continuous(
    name = "count",
    breaks = seq(0,40,2),
    labels = seq(0,40,2))
```

Histogram =



```
# Shape = Slightly Right Skewed
# Center =
mean(data_purchases$Purchases)
# Output = 7.005236 = 7 (approx)
median(data_purchases$Purchases)
# Output = 7
sd(data_purchases$Purchases)
# Output = 2.420303 = 2.4 (approx)

# Spread of data is
(7.005235-2.420303)
# Output = 4.584932 = 4.58 = 5
```

```
(7.005235+2.420303)
# Output = 9.425538 = 9.43 = 9
# The spread of data is between (5,9) in Purchases

getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

print(getmode(data_purchases$Purchases))
# The mode is 8

min(data_purchases$Purchases)
# Output = 2
max(data_purchases$Purchases)
# Output = 14
```

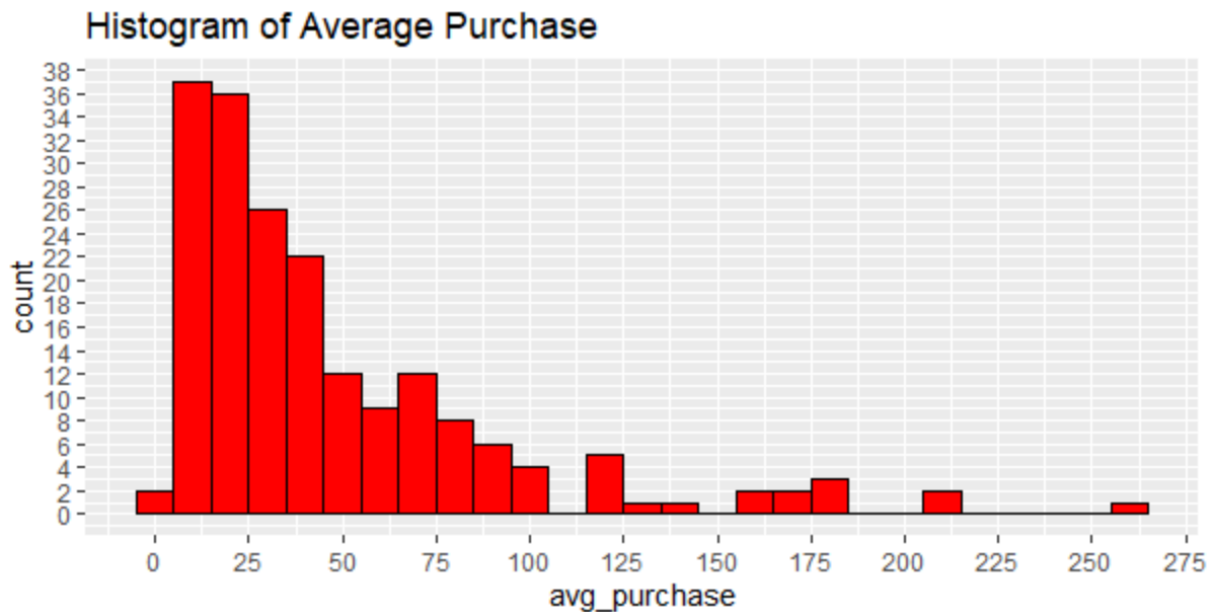
Insights from the Histogram of Purchases =

1. The number of times the frequent shopper card has been scanned for a purchase has been from 2 times to 14 times.
 2. You see more peaks around 7 and 8 which means that the greatest number of times the frequent shopper card was scanned is around 7 to 8 times.
 3. The bulk of times the shopper card used for purchases has been between 5 to 9.
 4. The mean is 7, median is 7 and standard deviation is 2.4.
-

Quantitative Variable - 3 = AvgPurchase

```
ggplot(  
  data = data_purchases,  
  aes(x = AvgPurchase)  
) +  
  geom_histogram(color = "black", fill = "red", binwidth = 10) +  
  ggtitle("Histogram of Average Purchase") +  
  scale_x_continuous(  
    name = "avg_purchase",  
    breaks = seq(0,280,25),  
    labels = seq(0,280,25)) +  
  scale_y_continuous(  
    name = "count",  
    breaks = seq(0,40,2),  
    labels = seq(0,40,2))
```

Histogram =



```
# Shape = Right Skewed Distribution  
# Center =  
mean(data_purchases$AvgPurchase)  
# Output = 47.58712 = 47.59 (approx)  
median(data_purchases$AvgPurchase)  
# Output = 32.35  
sd(data_purchases$AvgPurchase)  
# Output = 44.49005 = 44.50 (approx)  
# Spread of data is  
(47.59-44.50)  
# Output = 3.09  
(47.59+44.50)  
# Output = 92.09  
# The spread of data is between (3.09,92.09) in Average Purchases
```

```
min(data_purchases$AvgPurchase)
# Output = 2.38
max(data_purchases$AvgPurchase)
# Output = 256.42
```

Insights from the Histogram of Average Purchase =

1. The average amount that the patron has spent on purchases while using their frequent shopper card is between \$2.38 to \$256.42.
 2. You see more peaks around 10 and 20 which means that the greatest number of people spent around \$10 to \$20.
 3. There are some outliers meaning some person spent around \$255 and two persons around \$210 using the shoppers card.
 4. The mean is 47.59, median is 32.35 and standard deviation is 44.50.
 5. Most average purchases are from \$10 to \$50 with around 130 out of 191 people spending that amount using the shoppers card.
-

2. Exploratory Data Analysis

C. Create a bar chart for each of your categorical variables. Provide at least 3 insights that these graphs have provided you.

Answer to Question:2-C =

Qualitative Variable - 1 = Gender

```
ggplot(  
  data = data_purchases,  
  aes(x = Gender)  
) +  
  geom_bar(color = "black", fill = "red") +  
  xlab("gender") +  
  ggtitle("Bar Chart of Gender")
```

```
table(data_purchases$Gender)  
# Number of Females = 97  
# Number of Males = 94
```

Bar Graph =



Insights =

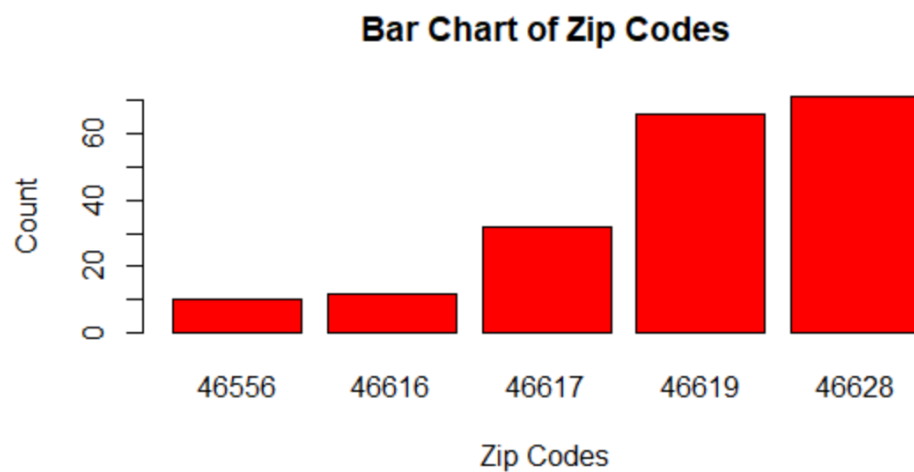
1. There are a greater number of female shoppers than male shoppers using the shoppers' card.

Qualitative Variable - 2 = Zip Code

```
barplot(table(data_purchases$Zip.Code),border = "black", col="red",  
        xlab = "Zip Codes", ylab = "Count",  
        main = "Bar Chart of Zip Codes")
```

```
table(data_purchases$Zip.Code)  
# Output =  
# 46556 46616 46617 46619 46628  
# 10    12    32    66    71  
  
length(data_purchases$Zip.Code)  
# Output = 191
```

Bar Chart =



Insights =

1. The 191 shoppers with shoppers' card in the data set are restricted to 5 zip codes which are close by to each other as they all start with 46 (same number).
 2. Customers from these two zip codes - 46619 and 46628 come to the store the most when purchasing with Shoppers Card.
-

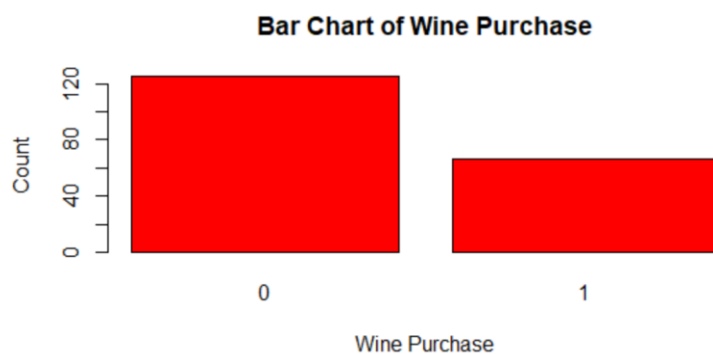
Qualitative Variable - 3 = Wine Purchases

```
barplot(table(data_purchases$WinePurchase),border = "black", col="red",
        xlab = "Wine Purchase", ylab = "Count",
        main = "Bar Chart of Wine Purchase")

table(data_purchases$WinePurchase)
# Output =
# 0 1
# 125 66

length(data_purchases$WinePurchase)
# Output = 191
```

Bar Chart =



Insights =

1. More shoppers with the shopper's card have not bought bottle of wine in their transactions where their frequent shopper card has been scanned.
-

2. Exploratory Data Analysis

D. Create a scatterplot with Purchases as the independent variable and Average Purchase as the dependent variable. Describe the relationship (Pattern, Direction and Strength) between these two variables. Provide the correlation.

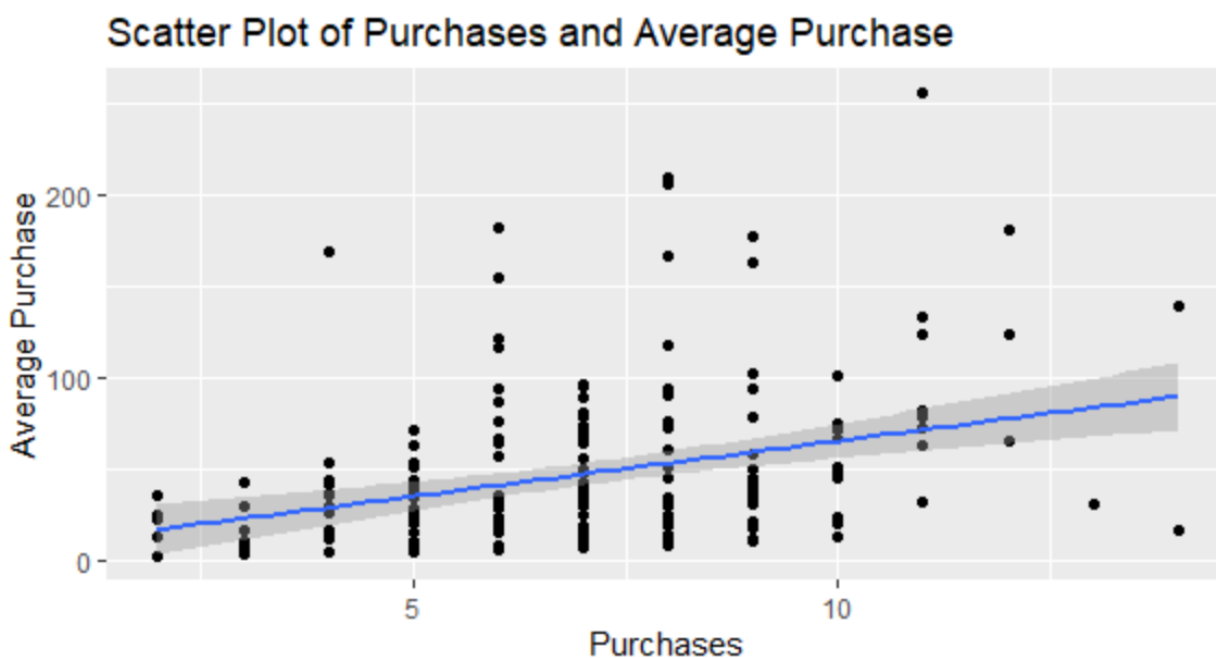
Answer to Question:2-D =

```
library(ggplot2)

ggplot(
  data = data_purchases,
  aes(x = Purchases,
      y = AvgPurchase)
) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = "lm") +
  xlab("Purchases") +
  ylab("Average Purchase") +
  ggtitle("Scatter Plot of Purchases and Average Purchase")

# Correlation Coefficient
cor(data_purchases$Purchases, data_purchases$AvgPurchase)
# Output = 0.3300256
```

Scatter Plot =



Pattern = The data set cluster around the straight line. Non-linear.

Direction = Positive Slope = Positive Relationship.

Strength = Purchases and Average Purchases are weak to moderately associated.

The Correlation Coefficient between Purchases and Average Purchases is 0.3300256.

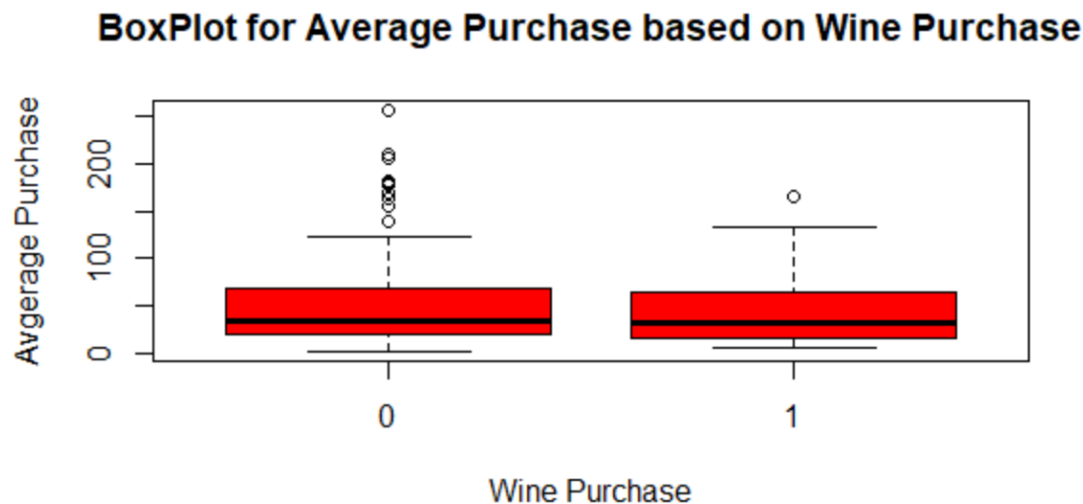
2. Exploratory Data Analysis

E. Create side by side box plots for Average Purchase, based on whether or not the patron has had a wine purchase. Comment on the relationship that you observe.

Answer to Question:2-E =

```
boxplot(data_purchases$AvgPurchase ~ data_purchases$WinePurchase ,  
col="red", main="BoxPlot for Average Purchase based on Wine Purchase",  
ylab="Average Purchase", xlab="Wine Purchase")
```

Box Plot =



Explanation =

The 0 indicates that the patron has not purchased a bottle of wine in any of the transactions where their frequent shopper card has been scanned and 1 indicates that the patron has purchased a bottle of wine in any of the transactions where their frequent shopper card has been scanned. So, the median for both 0 and 1 is almost the same and there are lot of outliers in 0 (when the wine purchase did not happen) indicating that people have spent money in other items in the store and whenever they spent large amount of money, they mostly did not buy Wine.

3. Modeling

A. The number of purchases is a count of how many purchases the patron has made and scanned their frequent shopper card. Propose a distribution for this data (including values for the parameter(s)). Show that your proposed distribution fits the given data.

Answer to Question:3-A =

The sample size is 191.

According to the Central Limit Theorem, when the sample size is large (> 30) the sample distribution of the sample mean will be Normal. I propose the Normal Distribution for Purchases. Since Purchases column values are quantitative and discrete, I propose a Normal Distribution with Continuity Correction.

Mean of Purchases = 7.005236

Standard Deviation = 2.420303

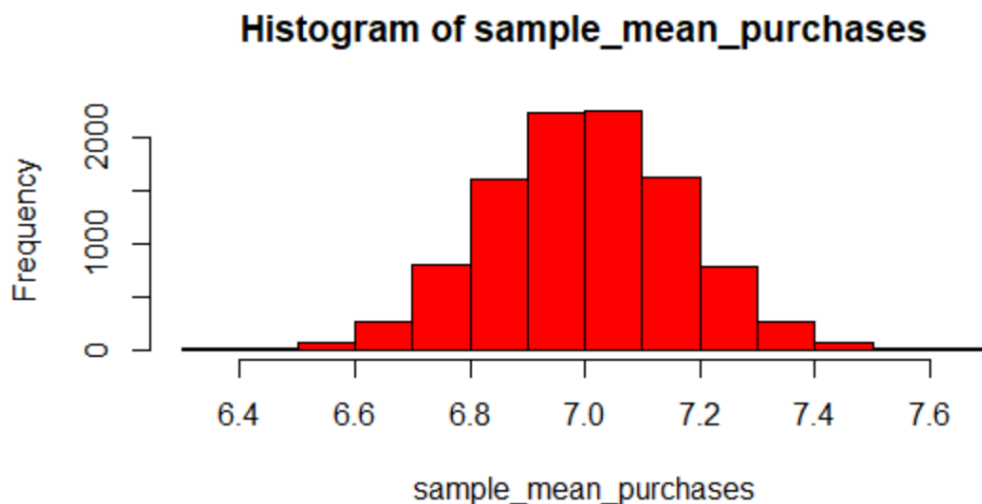
Purchases = $X = \text{Normal}(7.005236, 2.420303)$

The proposed distribution (Normal Distribution) fits the given data is by simulation of the sampling distribution of the sample mean as below which comes out to be Normal.

```
# Sampling Distribution of the Sample Mean
iters <- 10000
sample_mean_purchases <- matrix(NA,iters, 1)
for (i in 1:iters) {
  sample_purchases <- sample(data_purchases$Purchases, 100)
  sample_mean_purchases[i] <- mean(sample_purchases)
}

hist(sample_mean_purchases, col = 'red', borders ="black")
```

Histogram =



3. Modeling

B. If you divide the ages by the maximum age (78) you will get values between 0 and 1. Determine the values for α and β in the Beta distribution that provide a good model for these scaled ages. Show that your proposed distribution fits the given data.

Answer to Question:3-B =

```
ages_by_78 <- data_purchases$Age / 78
ages_by_78
```

Output =

```
[1] 0.6410256 0.6282051 1.0000000 0.6153846 0.2692308 0.7051282 0.6923077
[8] 0.6666667 0.7051282 0.8076923 0.5000000 0.3461538 0.9102564 0.8846154
[15] 0.2435897 0.6666667 0.8076923 0.5641026 0.7307692 0.4871795 0.2564103
[22] 0.8717949 0.5256410 0.8205128 0.8846154 0.6538462 0.4230769 0.7948718
[29] 0.2692308 0.2564103 0.9615385 0.7692308 0.9230769 0.4615385 0.4615385
```

```
mean(ages_by_78)
# Output = 0.6490804
var(ages_by_78)
# Output = 0.04145006
```

Since it is a beta distribution we have assumed =

$$E(X) = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$$

$$E(X) = \text{mean} = 0.65 \quad (\text{from the above})$$

$$\text{Var}(X) = \text{Variance} = 0.0415$$

$$\frac{\alpha}{\alpha + \beta} = 0.65 \Rightarrow \alpha = 0.65\alpha + 0.65\beta$$

$$\Rightarrow 0.35\alpha = 0.65\beta \Rightarrow \alpha = 1.86\beta \quad \text{--- ①}$$

$$\frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} = 0.0415 \Rightarrow \frac{(1.86\beta) \cdot \beta}{(\beta + 1.86\beta)^2 (\beta + 1.86\beta + 1)} = 0.0415$$

$$\Rightarrow \frac{1.86\beta^2}{(2.86\beta)^2 \cdot (2.86\beta + 1)} = 0.0415 \Rightarrow \frac{1.86\beta^2}{8.1786\beta^2} = (0.0415) \cdot (1 + 2.86\beta)$$

$$\Rightarrow 0.2274 = 0.0415 + 0.11869\beta$$

$$\Rightarrow 0.1859 = 0.11869\beta \Rightarrow \beta = \frac{0.1859}{0.11869} = 1.5663$$

$$\boxed{\beta = 1.5663}$$

$$\alpha = 1.86\beta = 2.9133 \Rightarrow \boxed{\alpha = 2.9133}$$

Next, we will calculate the PDF of Beta Distribution as below =

```
ages_by_78 <- data_purchases$Age / 78

alpha_value = 2.9133
beta_value = 1.5663
ages_by_78_probs <- pbeta(ages_by_78, alpha_value,beta_value)
ages_by_78_probs
```

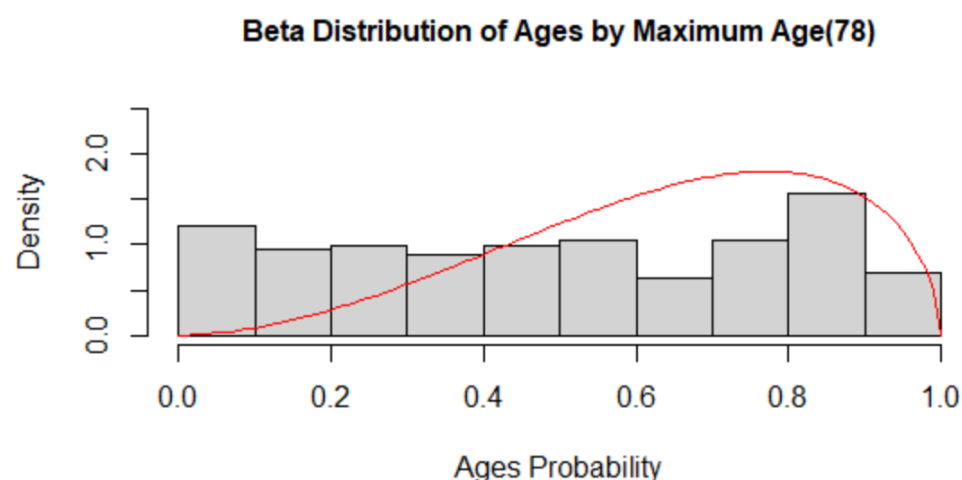
Output =

```
[1] 0.44410387 0.42324125 1.00000000 0.40278183 0.04540568 0.5535698
[7] 0.53108568 0.48694275 0.55356989 0.73814590 0.23978386 0.0904897
[13] 0.90972033 0.87033223 0.03438222 0.48694275 0.73814590 0.3253808
[19] 0.59918943 0.22426606 0.03965735 0.84948582 0.27244782 0.7610596
[25] 0.87033223 0.46534621 0.15503248 0.71506176 0.04540568 0.0396573
```

The proposed distribution is Beta Distribution and with the values of alpha and beta, we have calculated the probabilities of ages that were divided by 78 using the pbeta. Now we will plot the values of ages_by_78_probs using Histogram as below.

```
alpha_value = 2.9133
beta_value = 1.5663
hist(ages_by_78_probs, prob = T, main = "Beta Distribution of Ages by Maximum Age(78)",
     xlab = "Ages Probability", cex.main = 1.00, cex.lab = 1.00, cex.axis = 1.00,
     ylim = c(0,2.5))
x <- seq(min(ages_by_78_probs), max(ages_by_78_probs), 0.1)
curve(dbeta(x,alpha_value,beta_value), col = "red", add = T)
```

Histogram =



The above distribution is left skewed and represents the beta distribution because for beta distribution if alpha (2.9133 from above) is greater than beta (1.5663 from above), the distribution is left skewed. So, the proposed beta distribution fits the given data.

4. Inferential Statistics

For parts A through C:

- State the appropriate statistical test for the stated problem,
- Check conditions for inference for that test,
- Conduct the appropriate hypothesis test, and
- Interpret the results of the test.

A. In glancing at the graphs you provided the manager earlier, from your exploratory data analysis, she noticed that many of her customers come from the zip code 46628. She would like to know if the population of people from this zip code have an average total spending more than \$350. Hint: you will have to calculate total spending for patrons in order to do this problem.

Answer to Question:4-A =

```
library(tidyverse)
zip_code_46628_df <- data_purchases %>%
  filter(data_purchases$Zip.Code == '46628')

zip_code_46628_df[1:10,]
```

Output =

	Observation	Name	Gender	Age	Zip.Code	Purchases	AvgPurchase
1	1	Aaron Evans	Male	50	46628	5	8.10
2	2	Adam Davis	Male	49	46628	10	52.02
3	8	Andrew James	Male	52	46628	4	12.81
4	10	Anne Hoffman	Female	63	46628	5	9.36

```
no_of_people_in_46628 <- length(zip_code_46628_df$Observation)
no_of_people_in_46628
# Output = 71

zip_code_46628_df['TotalPurchase'] <- (zip_code_46628_df$Purchases*
  zip_code_46628_df$AvgPurchase)

avg_spending_of_people_in_46628 <- mean(zip_code_46628_df$TotalPurchase)
avg_spending_of_people_in_46628
# Output = 385.5187

sd_spending_of_people_in_46628 <- sd(zip_code_46628_df$TotalPurchase)
sd_spending_of_people_in_46628
# Output = 462.8603
```

1. We will be using t-distribution test for the problem

2. Conditions for inference for the test

1. $n > 30$ ($n = 71$)

2. Random sample

3. Population distribution is Normal

4. Population standard deviation is unknown

We will be using the t-distribution test as the appropriate statistical test for this problem.

3.

H_0 : Average total spending less than or equal to 350

H_a : Average total spending more than 350

```
t.test(zip_code_46628_df$TotalPurchase,mu = 350, alternative = "greater",
       conf.level = 0.95)
# Output =
# One Sample t-test

# data:  zip_code_46628_df$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
```

4.test-statistic = 0.6466

5.p-value = 0.26

6.p-value > alpha (0.05) so I fail to reject the Null Hypothesis.

7.Since we fail to reject the Null Hypothesis, we do not have enough evidence to believe that the average spending of customers from zip code 46628 is greater than 350.

4. Inferential Statistics

For parts A through C:

- State the appropriate statistical test for the stated problem,
- Check conditions for inference for that test,
- Conduct the appropriate hypothesis test, and
- Interpret the results of the test.

B. The store manager knows that the store is located on the border of zip codes 46556 (Notre Dame) and 46617 (a part of South Bend). Because of the convenience she knows that people are more likely to stop in regularly from these locations. She would like to know if the proportion of patrons that have at least 10 purchases is more for patrons from zip code 46556 than from zip code 46617.

Answer to Question:4-B =

1. We will use the **Two Sample Inference – Proportions Simulation** as our **Statistical Test** for this problem.

```
library(tidyverse)
zip_code_46556_df <- data_purchases %>%
  filter((Zip.Code == '46556'))
zip_code_46617_df <- data_purchases %>%
  filter((Zip.Code == '46617'))

no_of_people_in_46556 <- length(zip_code_46556_df$Observation)
no_of_people_in_46556
# Output = 10
no_of_people_in_46617 <- length(zip_code_46617_df$Observation)
no_of_people_in_46617
# Output = 32

zip_code_46556_df_10 <- zip_code_46556_df %>%
  filter(Purchases >= 10)
no_of_people_in_46556_10 <- length(zip_code_46556_df_10$Observation)
no_of_people_in_46556_10
# 2

zip_code_46617_df_10 <- zip_code_46617_df %>%
  filter(Purchases >= 10)
no_of_people_in_46556_10 <- length(zip_code_46617_df_10$Observation)
no_of_people_in_46556_10
# 4
```

2. All the Conditions for inference are not met as below.

- $np < 10 = 10 \cdot 2/10 < 10$
- $n(1-p) = 10 \cdot (8/10) < 10$
- $np < 10 = 32 \cdot 4/32 < 10$
- $n(1-p) = 32 \cdot (28/32) > 10$
- Random Sample was taken.

Since all the conditions for inference are not met, we use simulation.

3. State the Hypothesis =

3

$$H_0 : P_1 \leq P_2$$

$$H_a : P_1 > P_2$$

P_1 = Proportion of patrons that have atleast 10 purchases from
Zip Code 46556 (Notre Dame)

P_2 = Proportion of patrons that have atleast 10 purchases from
Zip Code 46617 (a part of South Bend)

```
prop.test(x = c(no_of_people_in_46556_10, no_of_people_in_46556_10),
          n = c(no_of_people_in_46556, no_of_people_in_46617),
          alternative = "greater")
# 2-sample test for equality of proportions with continuity correction

# data:  c(no_of_people_in_46556_10, no_of_people_in_46556_10) out of
c(no_of_people_in_46556, no_of_people_in_46617)
# X-squared = 2.1661, df = 1, p-value = 0.07054
# alternative hypothesis: greater
# 95 percent confidence interval:
# -0.06298598  1.00000000
# sample estimates:
# prop 1 prop 2
# 0.400  0.125

# Warning message:
# In prop.test(x = c(no_of_people_in_46556_10, no_of_people_in_46556_10), :
# Chi-squared approximation may be incorrect
```

4. Original test statistic =

```
origin.diff <- 2/10 - 4/32
origin.diff
# Output = 0.075
```

5. Simulation to calculate the p-value =

```
# Simulation
zip_code_46556 <- c(rep(1,2), rep(0,8))
zip_code_46556
zip_code_46617 <- c(rep(1,4), rep(0,28))
zip_code_46617
better <- c(zip_code_46556, zip_code_46617)
better
```

```

group <- c(rep("zip_code_46556",10), rep("zip_code_46617",32))
zip_code_data <- data.frame(cbind(better,group))

# 1. Find the original difference in proportions
origin.diff <- 2/10 - 4/32
origin.diff
# Output = 0.075

# 2. Number of Iterations
iters <- 10000

# 3. Initialize difference matrix
diff <- matrix(NA,iters,1)

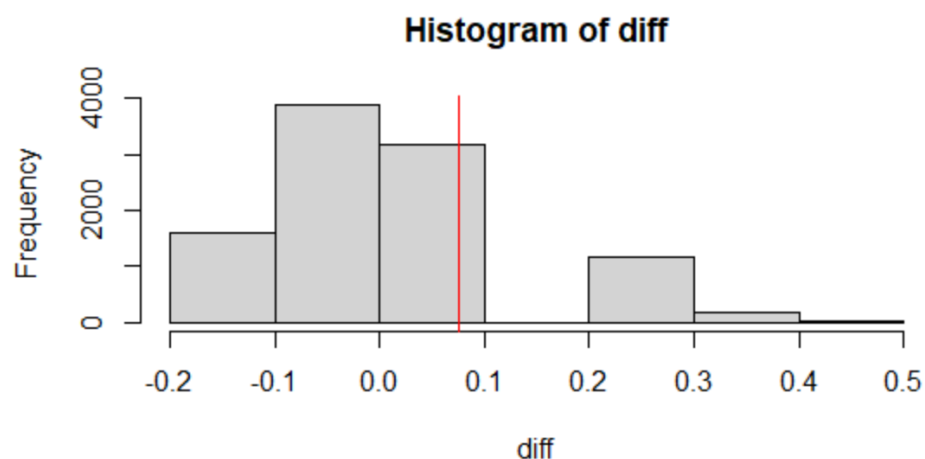
# 4. Run a loop to do this data collection 10000 times
temp <- group
for (i in 1:iters) {
  temp <- sample(temp, 42,replace=F)
  diff[i] <- sum(temp == "zip_code_46556" & better == 1)/10 -
    sum(temp == "zip_code_46617" & better == 1)/32
}

hist(diff,breaks=8)
abline(v = origin.diff,col = "red")

p_value <- sum(diff >= origin.diff)/iters
p_value
# Output = 0.4524

```

Histogram =



6. $p\text{-value} < \alpha$ (0.05) so I fail to reject the Null Hypothesis.

7. **Conclusion** = Since I fail to reject the Null Hypothesis. I do not have evidence to believe that the proportion of patrons that have at least 10 purchases is more for patrons from zip code 46556 than from zip code 46617.

4. Inferential Statistics

For parts A through C:

- State the appropriate statistical test for the stated problem,
- Check conditions for inference for that test,
- Conduct the appropriate hypothesis test, and
- Interpret the results of the test.

C. Patrons from zip code 46556 tend to be college students, whereas patrons from 46617 could be college students or local residents. Because of the difference in demographics, she would also like to know if there is a difference in the average amount spent between these two zip codes.

Answer to Question:4-C =

1. We will use the **Two Sample Statistical Inference – Independent Means** – Conducting an inference for a difference in Means as our Statistical Test for this problem.

```
zip_code_46556_df <- data_purchases %>%
  filter((Zip.Code == '46556'))
zip_code_46617_df <- data_purchases %>%
  filter((Zip.Code == '46617'))

no_of_people_in_46556 <- length(zip_code_46556_df$Observation)
no_of_people_in_46556
# Output = 10
no_of_people_in_46617 <- length(zip_code_46617_df$Observation)
no_of_people_in_46617
# Output = 32

zip_code_46556_df['TotalPurchase'] <- (zip_code_46556_df$Purchases*
  zip_code_46556_df$AvgPurchase)

avg_spending_of_people_in_46556 <- mean(zip_code_46556_df$TotalPurchase)
avg_spending_of_people_in_46556
# Output = 358.098
sd_spending_of_people_in_46556 <- sd(zip_code_46556_df$TotalPurchase)
sd_spending_of_people_in_46556
# Output = 238.6326

zip_code_46617_df['TotalPurchase'] <- (zip_code_46617_df$Purchases*
  zip_code_46617_df$AvgPurchase)

avg_spending_of_people_in_46617 <- mean(zip_code_46617_df$TotalPurchase)
avg_spending_of_people_in_46617
# Output = 244.6937
sd_spending_of_people_in_46617 <- sd(zip_code_46617_df$TotalPurchase)
sd_spending_of_people_in_46617
# Output = 230.7498
```

2. Conditions for inference =
 - Random Sample is taken.
 - The sample sizes are small (one of them is 10 as above).

```
var.test(zip_code_46556_df$TotalPurchase,zip_code_46617_df$TotalPurchase,
        alternative = "greater")
# Output =
# F test to compare two variances

# data:  zip_code_46556_df$TotalPurchase and zip_code_46617_df$TotalPurchase
# F = 1.0695, num df = 9, denom df = 31, p-value = 0.4116
# alternative hypothesis: true ratio of variances is greater than 1
# 95 percent confidence interval:
#  0.4862742      Inf
# sample estimates:
#  ratio of variances
# 1.06949
```

Since one of the sample sizes is 10 and is small, we need to assume the populations are close to normal. **Since the ratio of sample variance is almost equal to 1 (1.06949) as from the F-test above, we assume the population variances also need to equal.**

We also need to assume that $\sigma_1^2 = \sigma_2^2$

3. Hypothesis =

$$H_0 : \mu_{46556} - \mu_{46617} = 0$$

$$H_a : \mu_{46556} - \mu_{46617} \neq 0$$

μ_{46556} = Average amount spent by people from Zip code 46556

μ_{46617} = Average amount spent by people from Zip code 46617

4. We do a t-test

```
t.test(zip_code_46556_df$TotalPurchase,zip_code_46617_df$TotalPurchase,
       var.equal = TRUE)
# Two Sample t-test

# data:  zip_code_46556_df$TotalPurchase and zip_code_46617_df$TotalPurchase
# t = 1.3461, df = 40, p-value = 0.1859
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  -56.86745 283.67595
# sample estimates:
#  mean of x mean of y
# 358.0980 244.6937
```

5. test-statistic = 1.3461

6. p-value = 0.1859

7. $p\text{-value} > 0.05$, we fail to reject the Null Hypothesis.

8. Conclusion = Since we fail to reject the Null Hypothesis, we do not have evidence that there is a difference in the average amount spent between the two zip codes (46556, 46617).

4. Inferential Statistics

For parts A through C:

- State the appropriate statistical test for the stated problem,
- Check conditions for inference for that test,
- Conduct the appropriate hypothesis test, and
- Interpret the results of the test.

D. Barter Jacks is known for their inexpensive wines, which despite their price, make the store a lot of money. When students go home for the summer, wine sales normally drop. The manager would like to determine a marketing scheme for wine sales. In particular, she wants to know if she should target one gender or both with the marketing.

I. Determine the sample odds ratio for females that have purchased wine, versus males that have purchased wine. Interpret this value.

Answer to Question:4-D-I =

Observed Values =

		Gender		Total
		Female	Male	
Wine Purchase	1	53	13	66
	0	44	81	125
Total		97	94	191

$$\text{Odds ratio} = \frac{n_{11} * n_{22}}{n_{12} * n_{21}}$$

$$\text{Odds ratio} = (53*81)/(13*44) = 7.505$$

D. Barter Jacks is known for their inexpensive wines, which despite their price, make the store a lot of money. When students go home for the summer, wine sales normally drop. The manager would like to determine a marketing scheme for wine sales. In particular, she wants to know if she should target one gender or both with the marketing.

II. Create the confidence interval for the population odds ratio for females that have purchased wine versus males that have purchased wine. Interpret the results of this confidence interval.

Answer to Question:4-D-II =

Confidence Interval =

```
# sample_odds_ratio = (n11*n22)/(n12*n21)
sample_odds_ratio = (53*81)/(13*44)
sample_odds_ratio
# Output = 7.505245

# 95% Confidence Interval = log odds Confidence Interval
```



```

log_CI_1 = log(sample_odds_ratio) - 1.96*(sqrt(1/53 + 1/13 + 1/44 + 1/81))
log_CI_1
# Output = 1.30657
log_CI_2 = log(sample_odds_ratio) + 1.96*(sqrt(1/53 + 1/13 + 1/44 + 1/81))
log_CI_2
# Output = 2.724634

# 95% Confidence Interval = Original Confidence Interval = e to the power of log odds
CI_1 = exp(log_CI_1)
CI_1
# Output = 3.693482
CI_2 = exp(log_CI_2)
CI_2
# Output = 15.25084

# Since 1 does not fall in the interval (3.693482,15.25084), there is
# evidence of a relationship between the gender and purchase of wines.

```

D. Barter Jacks is known for their inexpensive wines, which despite their price, make the store a lot of money. When students go home for the summer, wine sales normally drop. The manager would like to determine a marketing scheme for wine sales. In particular, she wants to know if she should target one gender or both with the marketing.

III. Conduct the Chi-square test for independence between gender and having purchased wine and interpret your results

Answer to Question:4-D-III =

Observed Values =

		Gender		Total
		Female	Male	
Wine Purchase	1	53	13	66
	0	44	81	125
Total		97	94	191

Expected Values =

		Gender		Total
		Female	Male	
Wine Purchase	1	33.52	32.48	
	0	63.48	61.52	

```

observed <- matrix(c(53,13,44,81),nrow=2,ncol=2)
chisq.test(x = observed)

```

```
# Output =  
# Pearson's Chi-squared test with Yates' continuity correction  
  
# data: observed  
# X-squared = 33.375, df = 1, p-value = 7.601e-09
```

1.
Null Hypothesis = Gender and purchase of wines are Independent.
Alternative Hypothesis = Gender and purchase of wines are NOT Independent
2. test Statistic = 33.375
3. p-value = 7.601e-09
4. Since p-value < alpha (0.05) we reject the Null Hypothesis
5. Conclusion = Since we reject the Null Hypothesis, we do believe that there is a relationship between gender and purchase of wines and that Gender and Purchase of Wines are not independent.

Answer to Question-5 on next page.

5. Summary

Write a one to two page report to the manager describing insights from the data and suggestions from your tests that could help her with her goals. You may assume that the manager understands the basics of a hypothesis test, but cares more about answers than calculations. Make sure to address any concerns with data collection and how that could impact the results. Provide at least 2 suggestions for collecting better data for future analysis.

Answer to Question:5 =

Introduction = Barter Jacks – The Manager recently opened a store in South Bend, Indiana, close to Notre Dame's campus. The store is opened 3 months ago, and they have had 191 patrons sign up for their frequent shopper card. The data of 191 patrons is given to me to do the analysis and provide the insights of 191 patrons customer's shopping patterns to improve sales and bring more people into the store.

Analysis and Insights = After doing analysis based on the data on the 191 patrons who have shopper's card, I have come up with the following insights.

1. Though people from age 17 to age 78 have bought at the store using the shopper's card, I noticed that older people from age 60 to 68 have bought at the store more often. The most loyal customers have been between ages 35 to 67.
2. The most frequent number of times the shopper card was scanned was around 7 to 8 times which means they have been repeat customers and the average amount they spent was \$47.
3. Among Males and Females, there have been more females who purchased at the store.
4. People from zip codes of 46628 and 46619 happen to be the frequent purchasers at the store with the shopper's card but their average spending was not greater than \$350.
5. People who bought at the store mostly did not buy wine along with the other items they purchased.
6. There has been no strong evidence of people using the shopper's card more frequently and making large average purchases.
7. Women tend to purchase more wines when they come to shop with the shopper's card compared to males.

Suggestions to the Manager based on above Data Analysis and Insights = All these suggestions are based on the insights I drew from 191 patrons who used Shopper's Card. Middle age to older people are coming to the store more often to buy with the shopper's card. So, I would suggest the manager if he could do things to attract younger people and kids to come to the store. They have been many repeat customers, the manager needs to attract new customers by distributing the pamphlets about the store and its location and since it is close to Notre Dame's campus, he could do some marketing to attract University Students by displaying about the store in the Notice Board or other means of marketing. They have been more females to the store compared to the males so the manager can put and display more items for the male customers. People who generally bought at the store did not buy wine along with the purchase so the manager can put some discounts on the wine bottles to attract the customers to buy wine when they come to shop at the store. Women tend to buy wine more compared to men so the manager can add new stock of wines that the men like. Also, people generally buy around average of \$47 worth when they come to buy at the store. To increase the spending amount of the customers the manager can add more items to the store preferably for the college going people as the store is near the University.

Concerns with the data collection and its impact = The data is collected from those who bought at the store using the shopper's card, there could have been many people who came and bought at the store without the shopper's card, so we are losing large chunk of valid customers. Since the store is close to the Notre Dame's campus and the data shows that majority of the customers are in the middle age and older group, we don't know for sure if students bought at the store and not use the shopper's card or if they have been a smaller number of students who bought at the store.

Conclusion = I would advise the Manager that the insights from the data and the suggestions are strictly based on the people who used the shopper's card, and there could have been valid customers who could have bought at the store especially University of Notre Dame Students. So, to draw further insights from the data to improve sales and bring more people into the store, we need more data from people who did not use the shopper's card.

-----The End-----