** Project Report**
Name: Saloni Jain

Title

A Comparison of Online and Offline Prices in the Grocery Industry: A Case Study of Stop and Shop

Introduction

Stop & Shop is a major grocery store chain in Boston, offering a variety of products to its customers. With the rise of e-commerce, many consumers are opting for online shopping to save time and effort. However, it is essential to compare the prices of products in online and offline stores to determine if there is any significant difference. In this study, we will investigate if there is a statistically significant difference in prices for Stop and Shop products between their online and offline stores.

Hypothesis testing is a statistical technique used to determine whether a statement about a population parameter is true or not. It involves formulating a null hypothesis, which assumes that there is no difference between two populations, and an alternative hypothesis, which assumes that there is a difference between two populations (Cramer, D.,2003). In the sales industry, hypothesis testing can be used to determine if a particular sales strategy or promotion has a significant impact on sales. For example, a sales manager can use hypothesis testing to determine if there is a significant difference in sales between two different sales strategies (Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E.,2010).

Z-test, T-test, and F-test are statistical tests used to compare two populations. The Z-test is used when the sample size is large, and the population variance is known. The T-test is used when the sample size is small, and the population variance is unknown. The F-test is used to compare the variances of two populations (Field, A. P. ,2009). These tests can be applied in various fields such as quality control, education, and social sciences to compare the means of two groups and make statistical inferences (Jaccard, J., & Becker, M. A.,2002).

Referencing is an essential aspect of academic writing as it helps to acknowledge the sources of information used in the study. Proper referencing not only ensures that the work is credible but also helps to avoid plagiarism (Neville, C.,2007). It is essential to cite the sources of information used in the study accurately and consistently throughout the paper. Failure to reference properly can result in academic misconduct, which can affect the credibility of the research (Murray, R., & Hughes, I.,2008).

The dataset used in this report is M4Data_S_S.xlsx, which includes 55 randomly selected products from Stop & Shop store in Boston. The purpose of this dataset is to investigate if there is a statistically significant difference in prices between the online and offline stores for Stop and Shop products. The variables collected include Product Category, Product, Offline Price, and Online Price. These variables will help us to compare prices between online and offline stores for each product and product category and determine if there is a significant difference between them.

Analysis Section

Task 1.1

Problem and Research question

Describe the task: Clearly explain the problem and present the research question.

Problem:

```
knitr::include_graphics("~/Desktop/Northeastern/Courses/ALY6010_RProject/Images/StopnShop.jpeg")
```

The problem is that consumers may be unsure whether shopping online or in-store at Stop and Shop will yield better prices. This can make it difficult for them to make informed purchasing decisions

Therefore, it is important to investigate whether there is a significant difference in prices between the online and offline stores to provide consumers with the information they need to make informed decisions.

A dataset of 55 products has been collected, including their prices in both online and offline stores. The offline prices were collected from the Stop and Shop store located at 1620 Tremont Street in Boston.

The research question is:

Is there a statistically significant difference in prices for Stop and Shop products between their online and offline stores?"

Task 1.2

Variables

Describe the task: Explain the variables collected, and how do you believe they will help you answer the research question.

Do the Task

The variables collected for this study are:

Product Category: This variable refers to the general category of the product (e.g., dairy, produce, Bakery etc.). It will help us understand if there are any specific categories where there is a significant difference in prices between online and offline stores.

Product: This variable refers to the specific product being sold (e.g., milk,apple,oreo, etc.). It will help us understand if there are any specific products where there is a significant difference in prices between online and offline stores.

Offline Price: This variable refers to the price of the product in the offline store, measured in dollars. It will be used as a baseline to compare with the online prices and to calculate the difference between online and offline prices.

Online Price: This variable refers to the price of the product in the online store, measured in dollars. It will be used to compare with the offline prices and to calculate the difference between online and offline prices.

These variables will help us to analyze and answer the research question by allowing us to compare prices between online and offline stores for each product and product category, and determine if there is a significant difference between them.

Task 1.3

Present Table & Numerical variables graphs.

Describe the task: a) Present your data using a well-organized table using code kable(). b) present your numerical variables with minimum two different graphs and describe the data from each graph using minimum 3 different basic descriptive statistic definitions.

Part a (Present a table)

Do the Task

```
# Create table using kable() and wrap it with scroll_box()

M4_data = kable(M4Data_S_S, format = "html") %>%

  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F) %>%

  add_header_above(c("Table: Stop and Shop product prices in offline and online stores
" = 4)) %>%
```

```
  row_spec(0, bold = T, color = "white", background = "#6666FF") %>%
  column_spec(1:4, background = "#E0E0E0") %>%
  scroll_box(height = "400px", width = "100%")

# Print the formatted table to the console
M4_data
```

Table: Stop and Shop product prices in offline and online stores

Product Category	Product	Offline price (in dollars)	Online price (in dollars)
Dairy	Hood Whole Milk Gallon	8.18	7.09
Dairy	Hood Whole Milk Half Gallon	4.09	4.89
Dairy	SB Milk Whole Gallon	4.59	5.59
Dairy	SB Sliced Wite American Cheese	6.89	7.09
Dairy	Kraft Shredded Mozzarella cheese, 24oz	9.38	12.09
Dairy	Fage Total 0% Lowfat Yogurt	7.19	8.39
Dairy	Stonyfielt Organic Wholemilk Probiotic Yogurt Vanilla	5.49	6.49
Dairy	Land O Lakes Salted Butter	6.49	7.59
Dairy	Butter, The original	3.99	4.79
Dairy	SB Milk Whole Half Gallon	3.19	3.79
Produce	Red Bell Pepper each	1.69	2.21
Produce	Cucumber each	0.99	1.39
Produce	Yellow Onion each	0.84	1.11
Produce	Apple each	1.09	1.66
Produce	Roma Tomato each	0.69	0.55

Table: Stop and Shop product prices in offline and online stores

Product Category	Product	Offline price (in dollars)	Online price (in dollars)
Produce	Avocado each	1.25	2.09
Produce	Green Bell Pepper each	1.49	1.89
Produce	Russet Potato each	1.81	1.32
Produce	Lemon each	0.82	0.89
Produce	Cilantro	1.99	2.39
Produce	Shallot	0.69	0.45
Produce	Barlett peer	1.15	1.53
Produce	Cauliflower	2.50	5.09
Produce	Orange each	1.01	1.59
Snacks	Oreo Chocolate Cookies Family size	5.39	6.39
Snacks	Chips Ahoy Original, Party size	7.19	7.89
Snacks	Ritz Original Flavour Crackers, 1 box	4.66	5.09
Bakery	Nature’s Own Honey Wheat Bread	3.99	4.79
Bakery	Thomas Plain Bagels	5.29	6.49
Bakery	Nature’s Own 100% Whole Wheat Bread	3.99	4.79
Oils, Vinegars & Spices	Morton Iodized Salt	3.29	3.79
Oils, Vinegars & Spices	SB Vegetable Oil 64 FL Oz	7.29	8.49
Personal Care	Listerine Total Care Mouthwash, 1L	8.29	10.59

Table: Stop and Shop product prices in offline and online stores

Product Category	Product	Offline price (in dollars)	Online price (in dollars)
Personal Care	Cetaphil Moisturizing cream, 16oz	16.59	20.99
Personal Care	Dove Body wash, 22oz	8.49	9.89
Personal Care	Colgate Optic white advanced, 3.2 oz	5.69	7.59
Personal Care	Softsoap Antibacterial Liquid Hand Soap, 11.25 fl oz	3.29	3.79
Kitchen Supplies	Hefty Storage Slider Bags, 30x1	6.59	7.69
Beverages	Coca - Cola Mini Cola, 10x7.5 oz	7.99	9.39
Beverages	Gatorade Fruit Punch Flavoured Drink, 8x20 fl oz	10.79	12.69
Beverages	Starbucks breakfast blend medium roast ground, 12oz	10.99	12.89
Dry Goods	Goya Thai Jasmine Rice, 2lb	5.29	6.29
Dry Goods	Barilla classic blue box pasta spaghetti, 16oz	2.19	2.69
Dry Goods	Rao's Tomato basil Sauce, 24oz	7.99	9.39
Dry Goods	Goya Blackeye peas, 16oz	2.79	3.39
Dry Goods	Rao's Arrabbiata Sauce, 24oz	7.99	9.39
Dry Goods	Ragu classic alfredo sauce, 16oz	3.79	4.49
Breakfast	Skippy Peanut butter creamy, 40 oz	7.19	8.39
Breakfast	Quaker Fruit and cream instant oats, 8.4 oz	4.99	5.89
Breakfast	Smuckers Strawberry Jam, 18oz	4.99	5.89
Breakfast	Cheerios honey-nut cereal, 15.4oz	5.99	6.09

Table: Stop and Shop product prices in offline and online stores

Product Category	Product	Offline price (in dollars)	Online price (in dollars)
Breakfast	Nutella Chocolate Hazelnut Spread, 13oz	4.29	5.09
Breakfast	Kellogg's Corn flakes cereal, 18oz	5.69	9.29
Baking essentials	Hershey's Chocolate Syrup bottle, 24oz	3.99	4.79
Baking essentials	Betty Crocker Chocolate Fudge cake mix, 15.25oz	2.69	2.89

Observations

The M4Data table displays variables such as Product Category, Product, Offline price (in dollars), and Online price (in dollars) for various products available at both offline and online stores.

Although the products are the same in both types of stores, the prices differ significantly.

Upon analyzing the data, it can be observed that the majority of the products have a higher online price than their offline price. For instance, the product "SB Milk Whole Gallon" is priced at $4.59 in the offline Stop and Shop store, while it is priced at $5.59 in the online store.

This indicates that there is a difference in pricing for the same products in both types of stores.

To determine the significance of the price difference, further analysis is required.

Present Basic Descriptive Statistics for the M4 Dataset.

```
# use describe() to describe the M4 dataset
summary_m4 = describe(M4Data_S_S)


summary_final = round(summary_m4, 3)


# Create table using kable()
final_ds = kable(summary_final, format = "html") %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F) %>%
  add_header_above(c("Table: Basic Descriptive Statistics of Stop and Shop  product pr
ices in offline and online stores" = 14)) %>%
  row_spec(0, bold = T, color = "white", background = "#FF66B2") %>%
  column_spec(1:14, background = "#E0E0E0")
```

```
# Print
final_ds
```

Table: Basic Descriptive Statistics of Stop and Shop product prices in offline and online stores

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Product Category*	1	55	6.673	3.031	6.00	6.822	4.448	1.00	11.00	10.00	-0.155	-1.296	0.409
Product*	2	55	28.000	16.021	28.00	28.000	20.756	1.00	55.00	54.00	0.000	-1.266	2.160
Offline price (in dollars)	3	55	4.857	3.172	4.59	4.597	3.558	0.69	16.59	15.90	0.972	1.589	0.428
Online price (in dollars)	4	55	5.803	3.828	5.09	5.457	3.855	0.45	20.99	20.54	1.187	2.635	0.516

Observations

The table presents basic descriptive statistics for the Stop and Shop dataset.

The sample size n = 55.

The descriptive statistics provided suggest that the “Offline price” and “Online price” variables are continuous numerical variables, while “product” and “product category” variables are categorical variables.

The Offline price variable ranges from 0.69 to 16.59, with a mean of 4.857 and a standard deviation of 3.172.

The data is moderately skewed, with a skewness of 0.972, and has a kurtosis of 1.589, indicating that the data is slightly peaked.

The median value is 4.59, and the trimmed mean is 4.597.

The Online price variable ranges from 0.45 to 20.99, with a mean of 5.803 and a standard deviation of 3.828. The data is moderately skewed, with a skewness of 1.187, and has a kurtosis of 2.635, indicating that the data is very peaked.

The median value is 5.09, and the trimmed mean is 5.457.

Overall, the statistics suggest that there are differences in prices between offline and online stores, with higher prices for most products found in the online store. The data also indicates that there is variation in the pricing of products, as seen by the standard deviation values for the price variables.

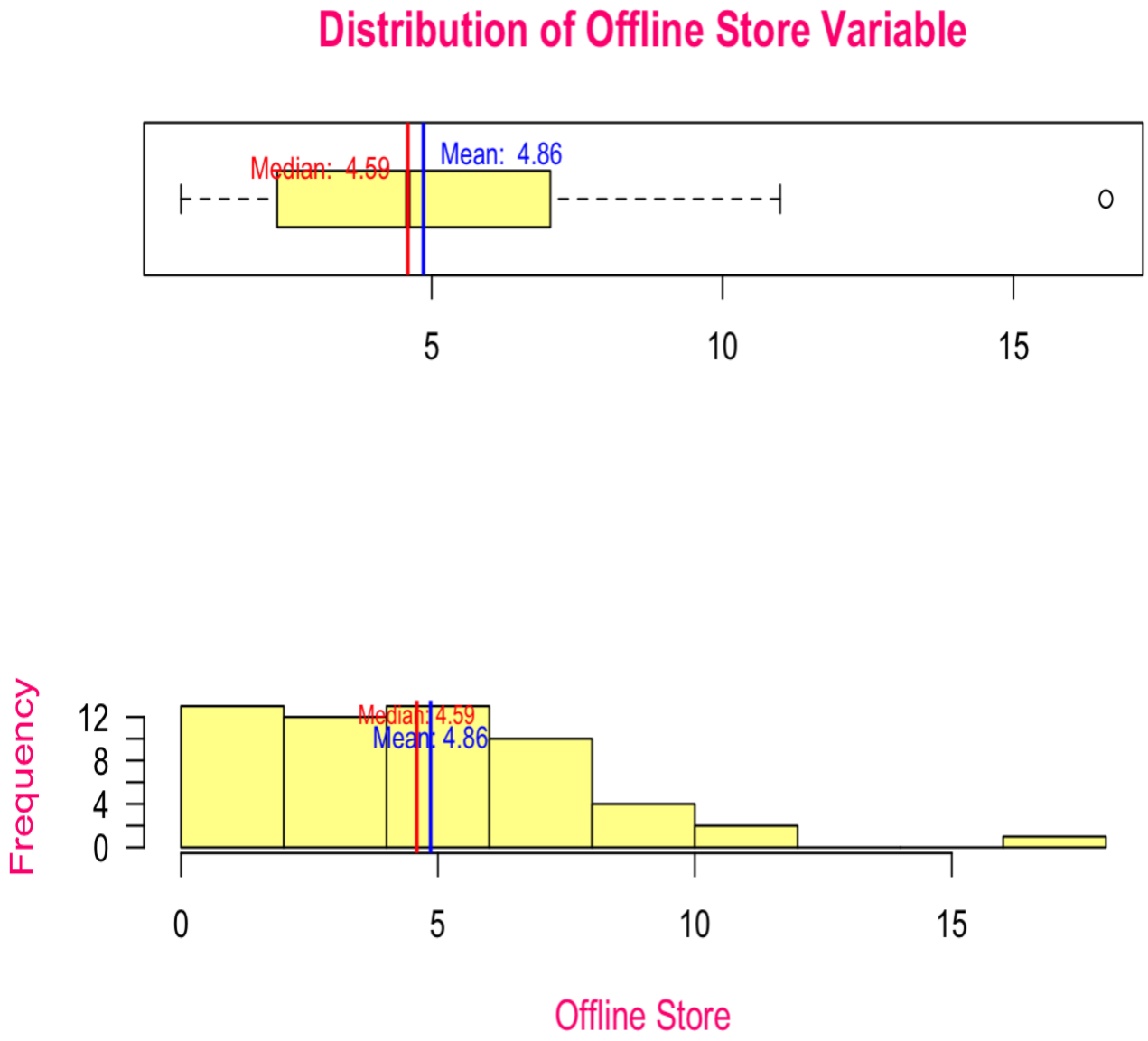
Part b (Present numerical variables using 2 different graphs)

Do the Task

Offline Store Data graphs:

```
# Set up plotting parameters
par(mfrow = c(2, 1))


# Create a horizontal box plot of the "Offline Store" variable
boxplot(x = M4Data_S_S$`Offline price (in dollars)`, horizontal = TRUE, main = "Distri
bution of Offline Store Variable",col.main ="#FF007F", col = "#FFFF99",  boxwex = 0.8,
        frame.plot = TRUE,
        col.lab = "#AE33B2",
        cex.main = 1.3,
        cex.lab = 0.8)



# Calculate the median and mean of the "Offline Store" variable
median_1 = median(M4Data_S_S$`Offline price (in dollars)`)
mean_1 = mean(M4Data_S_S$`Offline price (in dollars)`)

# Add the median and mean to the box plot
abline(v = median_1, col = "red", lwd = 2)
text((x = median_1),
      y = 1.2,
      labels = paste("Median: ", round(median_1, 2)),
      col = "red",
      pos = 2,
      cex = 0.8)


abline(v = mean_1, col = "blue", lwd = 2)
text((x = mean_1),
     y = 1.3,
     labels = paste("Mean: ", round(mean_1, 2)),
     col = "blue",
     pos = 4,
     cex = 0.8)


# Create a histogram of the "Offline store" variable
hist(M4Data_S_S$`Offline price (in dollars)`, main = NA, ylab = "Frequency", xlab = "O
ffline Store" ,col = "#FFFF99", boxwex = 0.5,
        frame.plot = TRUE,
```

```
        col.lab = "#FF007F",
        cex.lab = 1.1, las =1)

# Add the median and mean to the histogram
abline(v = median_1, col = "red", lwd = 2)
text((x=median_1),
     y = 15,
     paste("Median:",round(median_1, 2)),
     col = "red",
     cex = 0.7,
     pos = 1)


abline(v = mean_1, col = "blue", lwd = 2)
text((x=mean_1),
     y = 13,
     paste("Mean:",round(mean_1,2)),
     col = "blue",
     cex = 0.8,
     pos = 1)
```

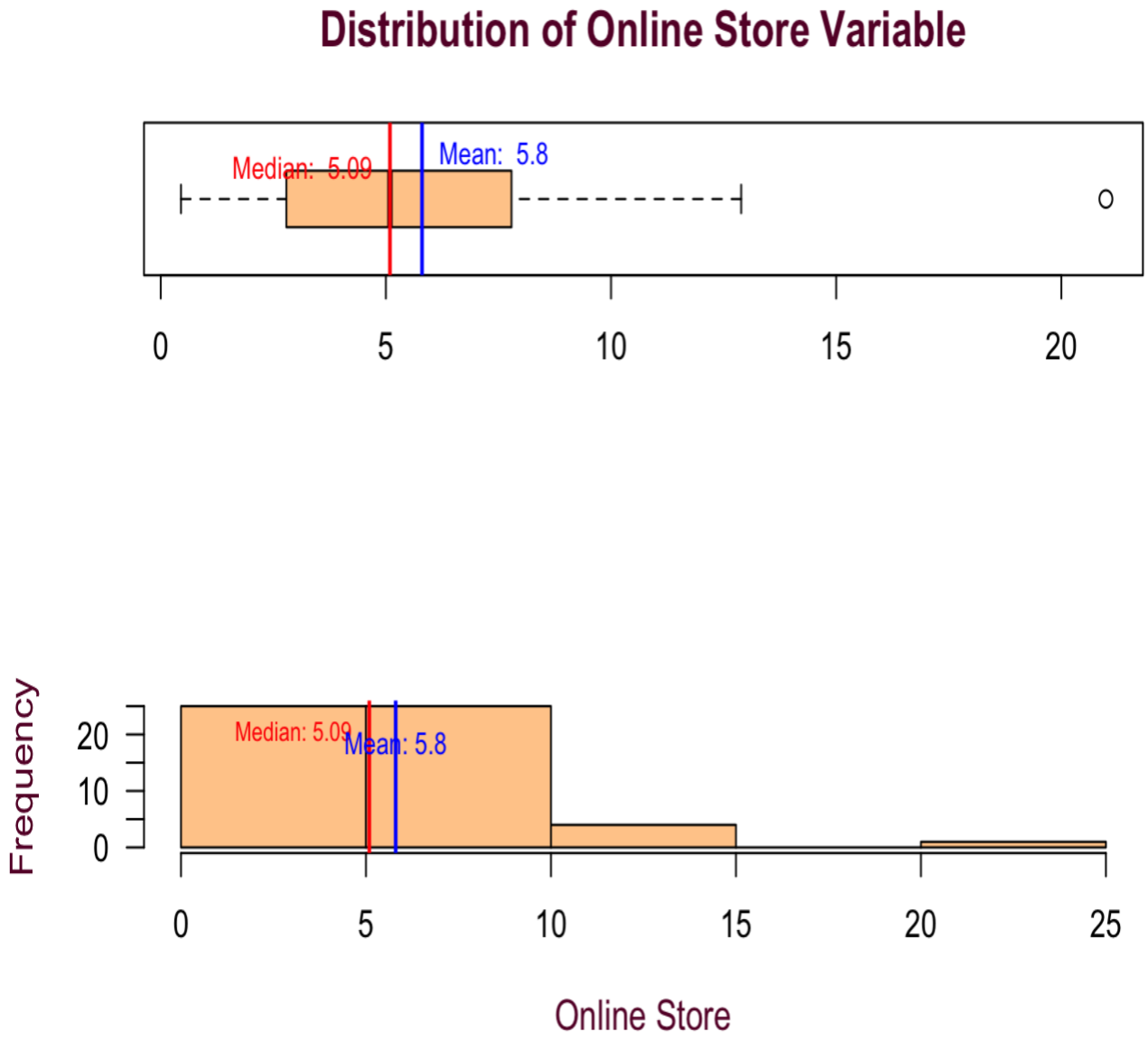


Online Store Data graphs:

```
# Set up plotting parameters

par(mfrow = c(2, 1))


# Create a horizontal box plot of the "Online Store" variable

boxplot(x = M4Data_S_S$`Online price (in dollars)`, horizontal = TRUE, main = "Distrib
ution of Online Store Variable",col.main ="#660033", col = "#FFCC99",  boxwex = 0.8,

        frame.plot = TRUE,

        col.lab = "#660033",
```

```
        cex.main = 1.3,
        cex.lab = 0.8)



# Calculate the median and mean of the "Online Store" variable
median_2 = median(M4Data_S_S$`Online price (in dollars)`)
mean_2 = mean(M4Data_S_S$`Online price (in dollars)`)

# Add the median and mean to the box plot
abline(v = median_2, col = "red", lwd = 2)
text((x = median_2),
      y = 1.2,
      labels = paste("Median: ", round(median_2, 2)),
      col = "red",
      pos = 2,
      cex = 0.8)

abline(v = mean_2, col = "blue", lwd = 2)
text((x = mean_2),
     y = 1.3,
     labels = paste("Mean: ", round(mean_2, 2)),
     col = "blue",
     pos = 4,
     cex = 0.8)

# Create a histogram of the "Online store" variable
hist(M4Data_S_S$`Online price (in dollars)`, main = NA, ylab = "Frequency", xlab = "On
line Store" ,col = "#FFCC99", boxwex = 0.5,
        frame.plot = TRUE,
        col.lab = "#660033",
        cex.lab = 1.1, las =1)

# Add the median and mean to the histogram
abline(v = median_2, col = "red", lwd = 2)
text((x=median_2),
     y = 20,
     paste("Median:",round(median_2, 2)),
```

```
     col = "red",
     cex = 0.7,
     pos = 2)



abline(v = mean_2, col = "blue", lwd = 2)
text((x=mean_2),
     y = 24,
     paste("Mean:",round(mean_2,2)),
     col = "blue",
     cex = 0.8,
     pos = 1)
```



Observations

For the offline store variable:

Boxplot:

The boxplot shows the distribution of the offline price variable, with a median of 4.59 and a mean of 4.857. The interquartile range (IQR), which is the range between the first and third quartiles, is between approximately 2.5 and 7.5. There is one outlier present above the upper whisker at 16.59, indicating that there is some variability in the pricing of products sold offline.

Histogram:

The histogram shows the distribution of the offline price variable, which is moderately skewed with a skewness of 0.972. The majority of the data is clustered between 0 and 10, with a peak at around 4.5. The histogram also confirms the presence of one outlier above 15, which is consistent with the boxplot.

Descriptive statistics:

The offline price variable has a range between 0.69 and 16.59, with a mean of 4.857 and a median of 4.59. The standard deviation is 3.172, indicating that there is some variability in the pricing of products sold offline. The data is moderately skewed, with a skewness of 0.972, and has a kurtosis of 1.589, indicating that the data is slightly peaked.

For the online store variable:

Boxplot:

The boxplot shows the distribution of the online price variable, with a median of 5.09 and a mean of 5.803. The IQR is between approximately 3 and 9.5. There is one outlier present above the upper whisker at 20.99, indicating that there is significant variability in the pricing of products sold online.

Histogram:

The histogram shows the distribution of the online price variable, which is highly skewed with a skewness of 1.187. The majority of the data is clustered between 0 and 10, with a peak at around 5. The histogram also confirms the presence of one outlier above 20, which is consistent with the boxplot.

Descriptive statistics:

The online price variable has a range between 0.45 and 20.99, with a mean of 5.803 and a median of 5.09. The standard deviation is 3.828, indicating that there is significant variability in the pricing of products sold online. The data is highly skewed, with a skewness of 1.187, and has a kurtosis of 2.635, indicating that the data is very peaked.

Task 1.4

Sampling method used & Indicate sample size.

Describe the task: a) describe the sampling method used, how, where the data was collected. b) Indicate your sample size c) Is your data described by the mean and standard deviation, or by proportion? Present values.

Do the Task

Part a

Sampling Method:

I collected the data from the online store of Stop and Shop using the Instacart app and from the physical store located at 1620 Tremont Street in Boston.

The products were chosen randomly from the product categories and from the population of Stop and Shop.

The data was collected on March 17th, 2023.

I did not ask anyone for help as the products were randomly chosen by me.

To make things simpler, I had written down the product full names from the online store while collected the online store data and its quantity, which helped me to map the product in the offline store.

Part b

Sample Size:

The sample size of the products is 55.

Part c

Data Description:

The data is described by both the mean and standard deviation as well as proportions.

The mean and standard deviation describe the prices of the products for online and offline stores. The mean for Offline store is `4.857` The mean for Online store is `5.803`

We can observe that the mean price for products in the online store of Stop and Shop is higher than that of the offline store.

The proportions is described by the frequency of each product category.

```
# present table for product category using kable()
table_categorical = as.data.frame(table(M4Data_S_S$`Product Category`))
colnames(table_categorical) = c("Product Category", "Frequency")

# arrange the table in descending order of frequency
table_categorical = arrange(table_categorical, desc(Frequency))
product_category_table = kable(table_categorical, format = "html") %>%

kable_styling(bootstrap_options = c("striped", "hover"), full_width = F) %>%
add_header_above(c("Table: Variable product category and its frequency" = 2)) %>%
row_spec(0, bold = T, color = "white", background = "#660066") %>%
column_spec(1:2, background = "#E0E0E0")

# print the table
product_category_table
```

Table: Variable product category and its frequency

Product Category	Frequency
Produce	14
Dairy	10
Breakfast	6
Dry Goods	6
Personal Care	5
Bakery	3
Beverages	3
Snacks	3
Baking essentials	2
Oils, Vinegars & Spices	2
Kitchen Supplies	1

Observations

The proportion of products in each category is also given, which can be useful for understanding the product distribution in Stop and Shop. We can observe that the majority of products belong to the produce and dairy categories with frequency of 14 & 10 respectively.

mentioning the proportions of each product category is still relevant as it provides additional information about the distribution of the sampled products across different categories. This can help to identify any potential biases in the sampling process and to gain insights into the product preferences of the sampled population.

Task 1.5

Null and Alternative hypotheses, Importance of presenting hypothesis.

Describe the task: a) State your null and alternative hypotheses. b) Explain the importance of well-presented hypotheses.

Part a

Based on the problem statement, the null and alternative hypotheses for this study are as follows:

Null hypothesis (Ho): There is no significant difference in prices for Stop and Shop products between their online and offline stores ($\mu 1 = \mu 2$).

Alternative hypothesis (Ha): There is a significant difference in prices for Stop and Shop products between their online and offline stores ($\mu 1 \neq \mu 2$).

Part b

A well-presented hypothesis helps to ensure that the study's results are reliable and valid. Without clear hypotheses, the study's findings may be difficult to interpret, and it may not be clear whether the results are statistically significant or due to chance. It provides a clear statement of the research question and help to ensure that the study is designed and executed in a way that allows for valid and reliable results.

The hypothesis is needed to determine whether there is a significant difference in prices between the online and offline stores for the same products. If there is a significant difference, it would provide important information for Stop and Shop to understand why there is a difference and how they can address it.

For example, if the online prices are significantly higher, they may need to adjust their pricing strategy or improve their online shopping experience to make it more attractive to customers. Conversely, if the offline prices are significantly higher, they may need to adjust their in-store pricing or promotions to remain competitive with other retailers.

Overall, a well-presented hypothesis helps to focus the study and provide meaningful insights for the company.

Task 1.6

Hypothesis is right-tailed, left-tailed, or two-tailed, confidence level and alpha (α) value, Critical values.

Describe the task: a) explain if your hypothesis is right-tailed, left-tailed, or two-tailed. b) Present the confidence level and alpha (α) value. c) Calculate the corresponding critical values (Z or T).

Do the Task

Part a

For the study, based on the problem statement we have stated the null hypothesis ($\mu 1 = \mu 2$) and the alternate hypothesis ($\mu 1 \neq \mu 2$).

There are three different types of hypothesis tests:

Two-tailed test: The alternative hypothesis contains the "≠" sign

Left-tailed test: The alternative hypothesis contains the "<" sign

Right-tailed test: The alternative hypothesis contains the ">" sign

Based on this data and our study we know that it is a two-tailed hypothesis test. As our alternate hypothesis ($\mu 1 \neq \mu 2$).

A two-tailed test was chosen because the alternative hypothesis states that there is a significant difference in prices between the two stores, but it does not specify the direction of the difference. Therefore, a two-tailed test is appropriate because it allows for the possibility that the mean price for the online store is either significantly higher or significantly lower than the mean price for the offline store.

Part b

A smaller sample size or a higher variability will result in a wider confidence interval with a larger margin of error. The level of confidence also affects the interval width. Therefore, if I want a higher level of confidence, that interval will not be as tight. A tight interval at 95% or higher confidence is ideal.

That is why for the further calculations I have choose confidence level of 95%.

In addition to the explanation above, the chosen confidence level of 95% means that if we were to repeat this study many times, we can be 95% confident that the true population mean difference between the prices of Stop and Shop products in their online and offline stores would fall within our calculated confidence interval. This level of confidence is commonly used in statistical analyses and provides a good balance between accuracy and precision.

The alpha (α) value is the significance level that we set for our test, which represents the probability of rejecting the null hypothesis when it is actually true.

```
# Setting the confidence level
ci_level = 0.95


sample_size_ss = nrow(M4Data_S_S)


df_ss = sample_size_ss - 1


alpha_ss = 0.05
```

The Confidence level is `0.95`.

The significance level (alpha) is `0.05`.

Degrees of freedom is `54`.

It is important to choose an appropriate alpha value because setting it too high can increase the risk of making a Type I error, while setting it too low can increase the risk of making a Type II error (failing to reject the null hypothesis when it is actually false). By setting a standard alpha value, we can compare our calculated p-value to this value to determine whether or not to reject the null hypothesis.

The appropriate statistical test for comparing prices of the same products sold online and offline at Stop and Shop is a paired samples t-test. This is because each product represents a pair of related observations. Additionally, we don't know the standard deviation of the population.

Part c

```
# calculate the corresponding critical values for t test
```

```
cv_left = qt(alpha_ss/2, df_ss)


cv_right = qt(1-alpha_ss/2, df_ss)
```

The critical values for alpha = 0.05 are `-2.005` and `2.005`.

Task 1.7

Normalized density distribution

Describe the task: a) Present the normalized density distribution with critical values, confidence level area, alpha area. b) Explain the data distribution using 3 different basic descriptive definition.

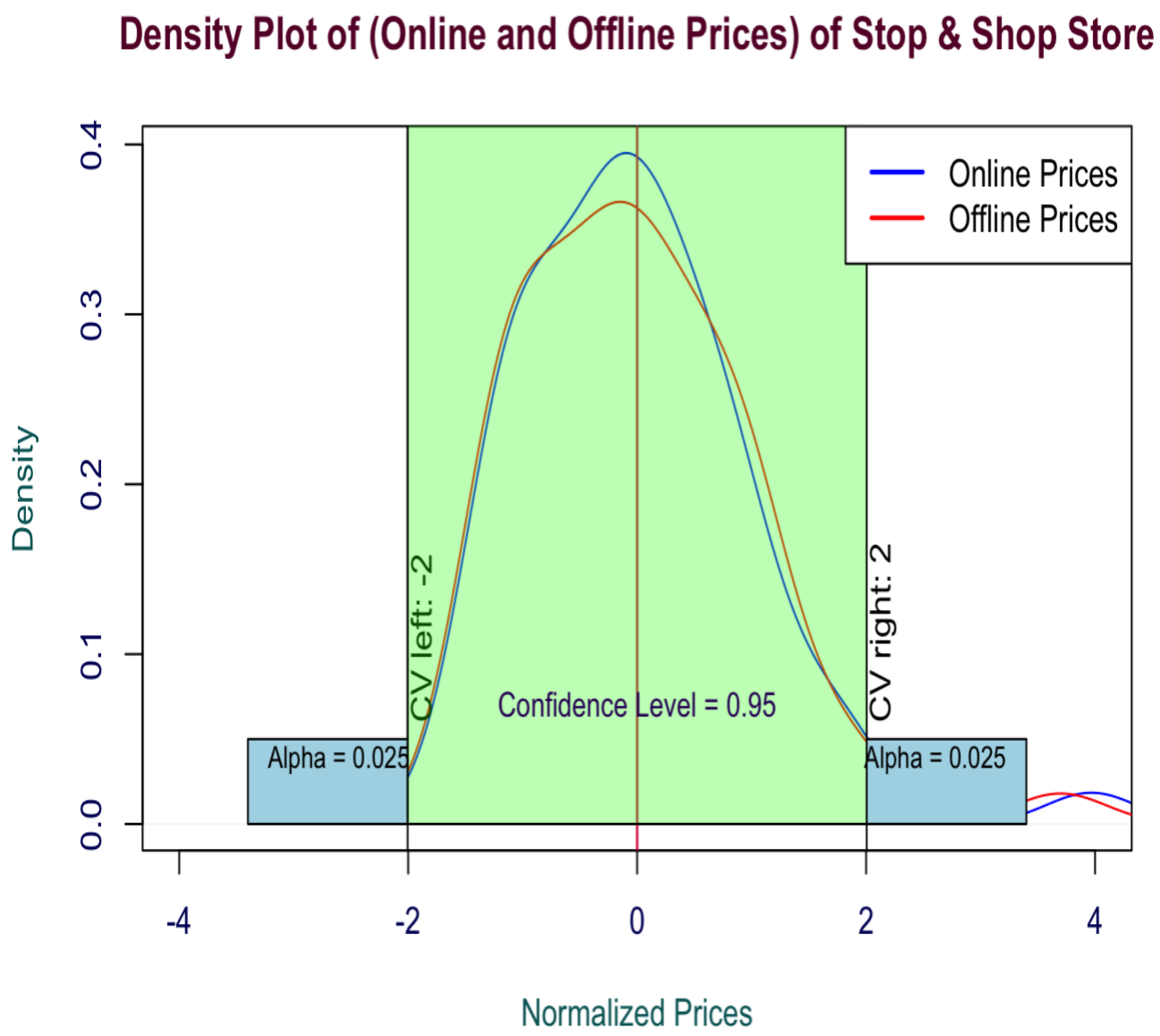
Do the Task

Part a:

```
offline_mean = mean(M4Data_S_S$`Offline price (in dollars)`)
offline_sd =  sd(M4Data_S_S$`Offline price (in dollars)`)


online_mean = mean(M4Data_S_S$`Online price (in dollars)`)
online_sd = sd(M4Data_S_S$`Online price (in dollars)`)


M4Data_S_S$Online_Normalized = (M4Data_S_S$`Online price (in dollars)` - online_mean)
/ online_sd
M4Data_S_S$Offline_Normalized = (M4Data_S_S$`Offline price (in dollars)` - offline_mea
n) / offline_sd


# Calculate mean and standard deviation of normalized data
online_mean_norm = mean(M4Data_S_S$Online_Normalized)
online_sd_norm = sd(M4Data_S_S$Online_Normalized)


offline_mean_norm = mean(M4Data_S_S$Offline_Normalized)
offline_sd_norm = sd(M4Data_S_S$Offline_Normalized)




# Set the plotting area
plot(density(M4Data_S_S$Online_Normalized), xlim=c(-4, 4), xlab="Normalized Prices", m
ain="Density Plot of (Online and Offline Prices) of Stop & Shop Store", col.main = "#6
60033", col.axis = "#000066", col.lab = "#006666", col = "blue")
lines(density(M4Data_S_S$Offline_Normalized), col="red")
```

```
# Add vertical lines for the means
abline(v = online_mean_norm, col = "blue")

abline(v = offline_mean_norm, col = "red")



# Calculate critical values for t-test and add vertical lines with labels
alpha = 0.05
df = 54
cv_left = qt(alpha/2, df)
cv_right = qt(1-alpha/2, df)

# cv_left
abline(v = cv_left)

text(x = cv_left, y = 0.06, paste("CV left:", round(cv_left, 2)), srt=90, pos=4)

rect(cv_left, 0.0, -3.4, 0.05, col = "lightblue")



#cv_right
abline(v = cv_right)

text(x = cv_right, y = 0.06, paste("CV right:", round(cv_right, 2)), srt=90, pos=4)

rect(cv_right, 0.0, 3.4, 0.05, col = "lightblue")

# Add a shaded area for the confidence level
rect(cv_left, 0, cv_right, 4, col = rgb(0, 1, 0, alpha = 0.3))

# Add labels for alpha and confidence level
text(x = (cv_left+cv_right)/2, y = 0.05, paste("Confidence Level =", 1-alpha), cex=0.9
, pos=3, col = "#330066")
text(x = cv_right+0.6, y = 0.02, paste("Alpha =", alpha/2), cex=0.8, pos=3)
text(x = cv_left-0.6, y = 0.02, paste("Alpha =", alpha/2), cex=0.8, pos=3)

# Add legend
```

```
legend("topright",
       legend = c("Online Prices", "Offline Prices"),
       col = c("blue", "red"),
       lty = c(1, 1),
       lwd = c(2, 2))
```



Observations

The R code performs the following steps:

Calculates the mean and standard deviation for the offline and online prices, Normalizes the data using the calculated mean and standard deviation, Calculates the mean and standard deviation for the normalized

data, Plots the density graph for both the normalized online and offline prices, Adds vertical lines for the means of the two distributions, Calculates the critical values for a t-test and adds vertical lines and labels for the left and right critical values, Adds a shaded area for the confidence level and labels for the confidence level and alpha, Adds a legend for the two distributions.

Plotted a density graph of the normalized data, with online prices in blue and offline prices in red.

The two graphs overlap each other, indicating that most of the product prices fall in the similar price range.

However, the density graph shows that the peak for the online store prices is higher than the peak for the offline store prices, suggesting that the online prices may be slightly higher on average.

The green shaded area represents the confidence level, which in this case is 0.95. This means that we can be 95% confident that the true difference between the online and offline prices falls within this area.

The alpha value represents the rejection region, which is the area beyond the critical values. In this case, the alpha value is 0.05, so the rejection region is the area beyond the critical values of -2.005 and 2.005.

Part b:

Skewness:

Skewness is a measure of the asymmetry of a probability distribution. If the skewness is negative, the distribution is skewed to the left (meaning the tail is longer on the left side). If the skewness is positive, the distribution is skewed to the right (meaning the tail is longer on the right side). By looking at the density plot in the code, we can see that the online prices and offline prices are skewed to the right.

Kurtosis:

Kurtosis is a measure of the "peakedness" of a probability distribution. A distribution with positive kurtosis has a sharper peak than the normal distribution, while a distribution with negative kurtosis has a flatter peak than the normal distribution. The kurtosis value for the online and offline prices in the descriptive statistics table is negative, which means the distributions are flatter than the normal distribution.

Range:

The range is the difference between the maximum and minimum values in a dataset. Looking at the descriptive statistics table, we can see that the range for both online and offline prices is relatively large, with a maximum value of 20.99 and 16.59, respectively, and a minimum value of 0.45 and 0.69, respectively. This suggests that there is a wide variation in prices for both online and offline stores.

Overall, the normalized density distribution and the descriptive statistics indicate that there is some overlap between the prices for online and offline stores, but there may be a slight difference in the average prices, with online prices being slightly higher on average.

Task 1.8

Test value

Describe the task: Calculate tTest value.

Do the Task

```
# Calculate the ttest value for the dependent sample


# Create a new column named "diff"
M4Data_S_S$diff = NA


# Loop through each row and calculate the difference of each product
for (i in 1:nrow(M4Data_S_S)) {
  M4Data_S_S$diff[i] = M4Data_S_S$`Offline price (in dollars)`[i] - M4Data_S_S$`Online price (in dollars)`[i]
}


diff_tot = sum(M4Data_S_S$diff)


# diff_tot


# Create a new column named "diff_square"
M4Data_S_S$diff_square = NA



# Loop through each row and calculate the square of each difference value
for (i in 1:nrow(M4Data_S_S)) {
  M4Data_S_S$diff_square[i] = M4Data_S_S$diff[i]^2
}


total_square_diff = sum(M4Data_S_S$diff_square)


# total_square_diff


# Find the mean of differences


D_bar =  diff_tot / sample_size_ss


# D_bar


# Find the standard deviation


top = (sample_size_ss * total_square_diff) - ((diff_tot)^2)
```

```
bottom = sample_size_ss * (sample_size_ss - 1)

sd_ttest = sqrt(top/bottom)

# sd_ttest

# calculate the ttest value for dependent samples
mu_d = 0 # as we had stated earlier that the null hypothesis is equal to 0
top1 = D_bar - mu_d

bottom1 = sd_ttest / sqrt((sample_size_ss))

ttest_dependent = top1 / bottom1

# ttest_dependent



task_8 = data.frame(
Difference = diff_tot,
Difference_Square = round(total_square_diff,3),
Mean_Difference = round(D_bar,3),
Standard_Deviation = round(sd_ttest,3),
tTest_value = round(ttest_dependent,3)

)

final_table = t(task_8 )

colnames(final_table) = c("Calculations")
```

Table: tTest Calculations for dependent two-sample hypothesis testing

	Calculations
Difference	-52.000
Difference_Square	94.542
Mean_Difference	-0.945
Standard_Deviation	0.917
tTest_value	-7.649

Observations

The tTest value for dependent two-sample hypothesis testing is -7.649. This value is obtained by calculating the difference, difference square, mean difference, and standard deviation of the data. The negative value of the mean difference (-0.945) indicates that on average, the online prices are higher than the offline prices.

The tTest value of -7.649 indicates that the difference between the means of the two samples is significant. This means that there is a significant difference between the online and offline prices. Since the tTest value is negative, we can infer that the online prices are significantly higher than the offline prices.

Task 1.9

Null hypothesis (true or false).

Describe the task: a) Depending on the direction of your test, ask the question: ttest value > CV or the question ttest value < CV, and explain the meaning of the TRUE or FALSE . b) Explain if your hypothesis testing analysis allowed you to reject your null hypothesis, or if you failed to reject your null hypothesis.

Do the Task

```
# Determine if null hypothesis is true or false
if (ttest_dependent <  cv_left) {
null_hypo_9 = "True"
} else {
null_hypo_9 = "False"
}



# null_hypo_9
```

Observations (Part a & b)

Since t value = `-7.649` is less than the critical value `-2.005` then:

Reject Ho = `True`

When conducting a hypothesis test using a t-test, we compare the calculated t-value with the critical values to determine whether we can reject the null hypothesis or not. If the calculated t-value falls outside of the range defined by the critical values, then we reject the null hypothesis, otherwise, we fail to reject it.

The alternative hypothesis is two-tailed ($\mu 1 \neq \mu 2$), so we ask the question “test > CV or test < CV.”

the tTest value `-7.649` is outside the critical value of -2.005 (Since the tTest value is negative), which are obtained for an alpha level of 0.05. This means that we can reject the null hypothesis and conclude that there is a significant difference between the online and offline prices at a 95% confidence level.

Therefore, based on the tTest calculations, we can say that there is evidence to support the claim that the online prices are significantly higher than the offline prices.

Therefore, we can reject the null hypothesis and conclude that there is a significant difference in prices for Stop and Shop products between their online and offline stores.

Task 1.10

P-value

Describe the task: Present P-value, explain the purpose and meaning of the P-value.

Do the Task

```
# Calculate p-value
pval_10 = pt(ttest_dependent, df_ss)



# pval_10
```

Observations

P-value = `1.8232558\times 10^{-10}`.

The P-value is a statistical measure that is used to determine the likelihood of obtaining the observed results of a hypothesis test, assuming that the null hypothesis is true.

In our case, we have conducted a dependent samples t-test to determine whether there is a significant difference in prices for Stop and Shop products between their online and offline stores. The null hypothesis states that there is no significant difference in prices between the two stores, while the alternative hypothesis states that there is a significant difference.

After conducting the t-test, we obtained a test statistic (t-value) of `-7.649` and a degrees of freedom of `54`. Using these values, we calculated the P-value, which was found to be `1.8232558\times 10^{-10}`.

Since this P-value is much smaller than the conventional level of significance ($\alpha = 0.05$) that we selected for our test, we can conclude that our results are statistically significant. Therefore, we can reject the null hypothesis and accept the alternative hypothesis, which states that there is a significant difference in prices for Stop and Shop products between their online and offline stores and we have enough evidence to support our decision to reject the null hypothesis.

Task 1.11

Normalized density distribution updated.

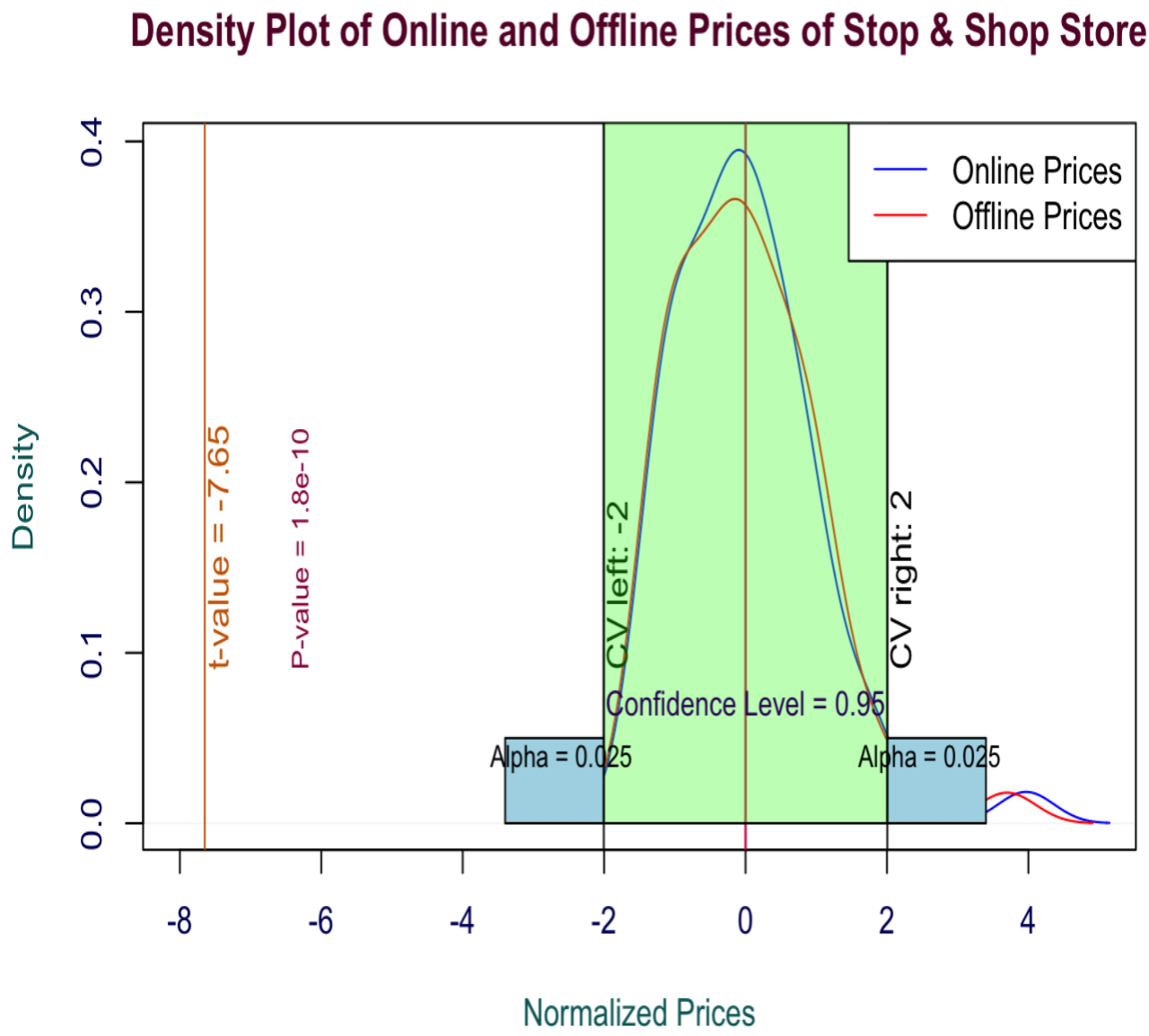
Describe the task: Present the normalized density distribution with critical values, confidence level area, alpha area, ttest value, p-value and write if that position matches the answer for True or False given in task 1.9.

Do the Task

```
offline_mean = mean(M4Data_S_S$`Offline price (in dollars)`)

offline_sd =  sd(M4Data_S_S$`Offline price (in dollars)`)


online_mean = mean(M4Data_S_S$`Online price (in dollars)`)

online_sd = sd(M4Data_S_S$`Online price (in dollars)`)


M4Data_S_S$Online_Normalized = (M4Data_S_S$`Online price (in dollars)` - online_mean)
/ online_sd

M4Data_S_S$Offline_Normalized = (M4Data_S_S$`Offline price (in dollars)` - offline_mea
n) / offline_sd


# Calculate mean and standard deviation of normalized data

online_mean_norm = mean(M4Data_S_S$Online_Normalized)

online_sd_norm = sd(M4Data_S_S$Online_Normalized)


offline_mean_norm = mean(M4Data_S_S$Offline_Normalized)

offline_sd_norm = sd(M4Data_S_S$Offline_Normalized)




# Set the plotting area

plot(density(M4Data_S_S$Online_Normalized), xlim=c(-8, 5), xlab="Normalized Prices", m
ain="Density Plot of Online and Offline Prices of Stop & Shop Store", col.main = "#660
033", col.axis = "#000066", col.lab = "#006666", col = "blue")

lines(density(M4Data_S_S$Offline_Normalized), col="red")


# Add vertical lines for the means

abline(v = online_mean_norm, col = "blue")


abline(v = offline_mean_norm, col = "red")




# Calculate critical values for t-test and add vertical lines with labels

alpha = 0.05

df = 54

cv_left = qt(alpha/2, df)

cv_right = qt(1-alpha/2, df)
```

```
# cv_left
abline(v = cv_left)


text(x = cv_left, y = 0.09, paste("CV left:", round(cv_left, 2)), srt=90, pos=4)


rect(cv_left, 0.0, -3.4, 0.05, col = "lightblue")



#cv_right
abline(v = cv_right)


text(x = cv_right, y = 0.09, paste("CV right:", round(cv_right, 2)), srt=90, pos=4)


rect(cv_right, 0.0, 3.4, 0.05, col = "lightblue")


# Add a shaded area for the confidence level
rect(cv_left, 0, cv_right, 4, col = rgb(0, 1, 0, alpha = 0.3))


# Add labels for alpha and confidence level
text(x = (cv_left+cv_right)/2, y = 0.05, paste("Confidence Level =", 1-alpha), cex=0.9
, pos=3, col = "#330066")
text(x = cv_right+0.6, y = 0.02, paste("Alpha =", alpha/2), cex=0.8, pos=3)
text(x = cv_left-0.6, y = 0.02, paste("Alpha =", alpha/2), cex=0.8, pos=3)




# add tTest value to the plot
abline(v = ttest_dependent, col = "#cc6600")
text(x = ttest_dependent, y = 0.09, labels = paste0("t-value = ", round(ttest_dependen
t, 2)),srt = 90, pos=4, col = "#cc6600")


#add p-value to the plot


text(x = -6.5, y = 0.09, labels = paste0("P-value = ", format(pval_10, digits = 2, sci
entific = TRUE)), cex=0.8,srt = 90, pos=4, col = "#99004c")



# Add legend
```

```
legend("topright",
       legend = c("Online Prices", "Offline Prices"),
       col = c("blue", "red"),
       lty = c(1, 1),
       lwd = c(1, 1))
```



Observations

Earlier, we calculated the t-value and p-value and compared them with the critical value for the dependent samples t-test.

After plotting the density plot of the normalized data for both online and offline prices, it can be observed graphically that there is a significant difference between the two groups. The t-value is negative and significant at -7.65, and the p-value is very small (1.8232558e-10), indicating strong evidence against the null hypothesis.

The critical value for a two-tailed test with a 5% significance level and 54 degrees of freedom is approximately -2. We can see that the test value falls well beyond this critical value, providing additional support for rejecting the null hypothesis.

Overall, both the statistical test and the graphical representation suggest that there is a significant difference between the online and offline prices of the Stop & Shop store.

Task 1.12

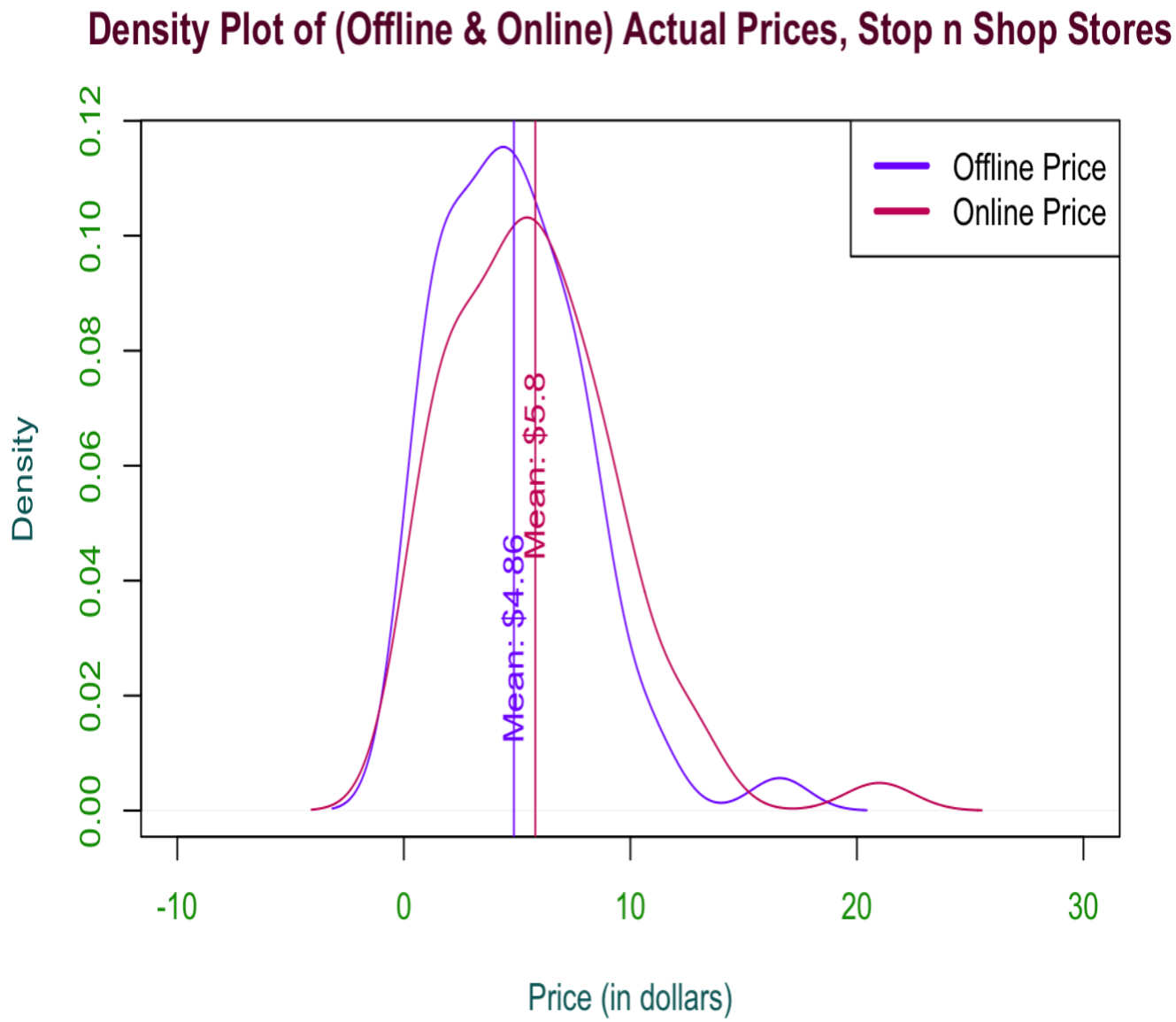
Density plot with the actual data.

Describe the task: Create a density plot with the actual data (offline & online prices), add meaningful observations, write 3 different basic descriptive statistic definitions.

Do the Task

```
# Set the plotting area

plot(density(M4Data_S_S$`Offline price (in dollars)`), xlim=c(-10, 30), xlab="Price (i
n dollars)", main="Density Plot of (Offline & Online) Actual Prices, Stop n Shop Store
s", col.main = "#660033", col.axis = "#009900", col.lab = "#006666", col = "#7F00FF")


# Overlay the density line for the second sample with a different color

lines(density(M4Data_S_S$`Online price (in dollars)`), col="#CC0066")


# Add vertical lines for the means

abline(v = mean(M4Data_S_S$`Offline price (in dollars)`), col = "#7F00FF")

abline(v = mean(M4Data_S_S$`Online price (in dollars)`), col = "#CC0066")


# Add text for both means

text(mean(M4Data_S_S$`Offline price (in dollars)`), 0.03, paste0("Mean: $", round(mean
(M4Data_S_S$`Offline price (in dollars)`), 2)), col = "#7F00FF", srt = 90)

text(mean(M4Data_S_S$`Online price (in dollars)`), 0.06, paste0("Mean: $", round(mean(
M4Data_S_S$`Online price (in dollars)`), 2)), col = "#CC0066", srt = 90)




# Add a legend

legend("topright",

       legend = c("Offline Price", "Online Price"),
```

```
        col = c("#7F00FF", "#CC0066"),
        lty = c(1, 1),
        lwd = c(3, 3))
```



Observations

The density plot depicts the distribution of actual prices of a product in two types of stores: offline and online. The plot shows that both offline and online prices are right-tailed, with a small peak for the maximum values.

The offline price distribution has a higher peak than the online price distribution, indicating that the majority of offline prices are clustered in a smaller range compared to the online prices.

The mean of offline prices is 4.86, while the mean of online prices is 5.8. This indicates that the prices of the product in the online store are relatively higher than in the offline store.

Descriptive statistics are used to summarize and describe the characteristics of a dataset.

Three basic descriptive statistics for this dataset are:

Mean: The mean is the average value of the prices in the dataset. It is calculated by summing all the values and dividing the sum by the total number of values. The mean of the offline store prices is $4.86, while the mean of the online store prices is $5.8. This indicates that the online store prices are higher on average compared to the offline store prices.

Standard Deviation: The standard deviation measures the spread of the dataset around the mean. It is calculated by taking the square root of the variance, which is the average of the squared differences from the mean.In this dataset, the standard deviation for the offline store prices is 3.17, while the standard deviation for the online store prices is 3.82. This indicates that the online store prices have a greater spread compared to the offline store prices.

Skewness: Skewness measures the degree of asymmetry in the dataset. A positive skewness value indicates that the tail of the distribution is longer on the positive side, while a negative skewness value indicates that the tail of the distribution is longer on the negative side. In this dataset, both the online and offline store prices have a right-skewed distribution, as indicated by the small peak for the maximum values and the longer right tail in the density plot.

Conclusions

In conclusion, this project involved analyzing and interpreting data from two different stores, an offline store and an online store. The aim was to compare the prices of products sold in both stores and determine whether there was a significant difference between the two.

Based on the results of the analysis, it can be concluded that there is a significant difference in prices between the online and offline stores of Stop and Shop. The offline store tends to sell products at lower prices than the online store. This information can be useful for consumers who are deciding whether to shop online or in-store, as it can help them make more informed purchasing decisions.

It is also recommended that consumers compare prices between the online and offline stores of Stop and Shop to make more informed purchasing decisions. Additionally, Stop and Shop could consider adjusting their pricing strategies to make their online and offline prices more competitive and consistent.

In terms of new skills gained, this project provided an opportunity to practice data analysis, data visualization, and statistical hypothesis testing. It also involved critical thinking, problem-solving, and careful data preparation and processing to ensure accurate results. Specifically, we learned about two-sample hypothesis testing, how to normalize data, and how to determine the null and alternative hypotheses. These skills can be valuable in a wide range of applications in both academic and professional settings.

Bibliography

1. Cramer, D. (2003). Advanced quantitative data analysis. Open University Press.

2. Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis (7th ed.). Prentice Hall.

3. Field, A. P. (2009). Discovering statistics using SPSS (3rd ed.). Sage Publications.

4. Jaccard, J., & Becker, M. A. (2002). Statistics for the behavioral sciences (4th ed.). Duxbury.

5. Neville, C. (2007). The complete guide to referencing and avoiding plagiarism (2nd ed.). Open University Press.

6. Murray, R., & Hughes, I. (2008). Writing up qualitative research (3rd ed.). Sage Publications.