MATH1324 Assignment 2

Code

Supermarket Price Wars

Group/Individual Details

- Gagan Deep Mullagur (s3804055)
- Jeyakaran Karnan (s3773303)
- Gayathri Jayabal (s3805886)

Executive Statement

The report contains a statistical analysis of the prices between Coles and Woolworths, two of Australia's leading supermarket chains. The analysis is to assess which supermarket is cheaper than the other based on the summary statistics of the prices of products available at both the supermarkets.

Online survey was conducted to collect the data required for the study from the respective websites of each supermarket. The dataset contains four attributes which are Products, Coles Price, Woolworths Price and Category. An extra factor variable called store is later created for grouping the observations according to their respective supermarkets. Summary statistics have been presented as part of descriptive statistics and a two-sample t-test for independent variables was carried out for statistical inference.

As a part of descriptive statistics, mean, median, mode, 1st quartile, and other features have been explicated. Boxplot for both Coles and Woolworths are used to visualize the statistics. The normality of the data is visualized using q-q plots and we could see that the majority of the observations lay inside the confidence interval. However, we considered CLT to assume normality of the distribution as the sample size was 180(n>30).

Load Packages and Data

Hide

```
#Importing Libraries
library(readr)
library(rmarkdown)
library(readxl)
library(magrittr)
library(dplyr)
library(car)
library(kableExtra)
#Reading Excel File
PriceWars <- read_csv("C:/Users/HP/Downloads/PriceWars.csv")
```

```
Parsed with column specification:
cols(
  Sno = □[32mcol_double()□[39m,
  Products = □[31mcol_character()□[39m,
  `Coles Price` = □[32mcol_double()□[39m,
  `Woolworth Price` = □[32mcol_double()□[39m,
  Category = □[31mcol_character()□[39m
)
```

Summary Statistics

It can be observed from the boxplots of both Coles and Woolworths prices that there are very few outliers. This may be the result of considering random samples of the data. Also, means of Coles data seem to be almost same when compared to Woolworths.

When we see the plot for all the products we notice that there are no significant price differences when we compare both the supermarkets. Although the plots of both supermarkets are almost similar, we can see that median price for Coles is almost equal to that of Woolworths which implements that both supermarket prices does not have much difference. Also, we can see few outliers because of taking random variables from different categories.

Hide

```
ColesPrices <- PriceWars %>%
  summarise(Observations = n()
            , Mean = mean(PriceWars$`Coles Price`)
            ,Median = median(PriceWars$`Coles Price`)
            ,'Standard Deviation' = sd(PriceWars$`Coles Price`)
            ,'First Quantile' = quantile(PriceWars$`Coles Price`, .25)
            ,'Third Quantile' = quantile(PriceWars$`Coles Price`, .75)
            ,'Inter Quantile' = quantile(PriceWars$`Coles Price`, .75) - quantile(PriceWars$`C
oles Price`, .25)
            , Minimum = min(PriceWars$`Coles Price`)
            , Maximum = max(PriceWars$`Coles Price`)
            , Missing = sum(is.na(PriceWars$`Coles Price`)))

WoolworthsPrices <- PriceWars %>%
  summarise(Observations = n()
            , Mean = mean(PriceWars$`Woolworth Price`)
            ,Median = median(PriceWars$`Woolworth Price`)
            ,'Standard Deviation' = sd(PriceWars$`Woolworth Price`)
            ,'First Quantile' = quantile(PriceWars$`Woolworth Price`, .25)
            ,'Third Quantile' = quantile(PriceWars$`Woolworth Price`, .75)
            ,'Inter Quantile' = quantile(PriceWars$`Woolworth Price`, .75) - quantile(PriceWar
s$`Woolworth Price`,.25)
            , Minimum = min(PriceWars$`Woolworth Price`)
            , Maximum = max(PriceWars$`Woolworth Price`)
            , Missing = sum(is.na(PriceWars$`Woolworth Price`)))

kable(ColesPrices, caption = 'Summary Statistics for Coles Prices') %>% kable_styling(bootstra
p_options = c("hover","condensed"))
```

Summary Statistics for Coles Prices

Observations	Mean	Median	Standard Deviation	First Quantile	Third Quantile	Inter Quantile	Minimum	Maximum	Missing
180	4.056278	3.675	2.124828	2.5	5.3075	2.8075	0.33	11	0

Hide

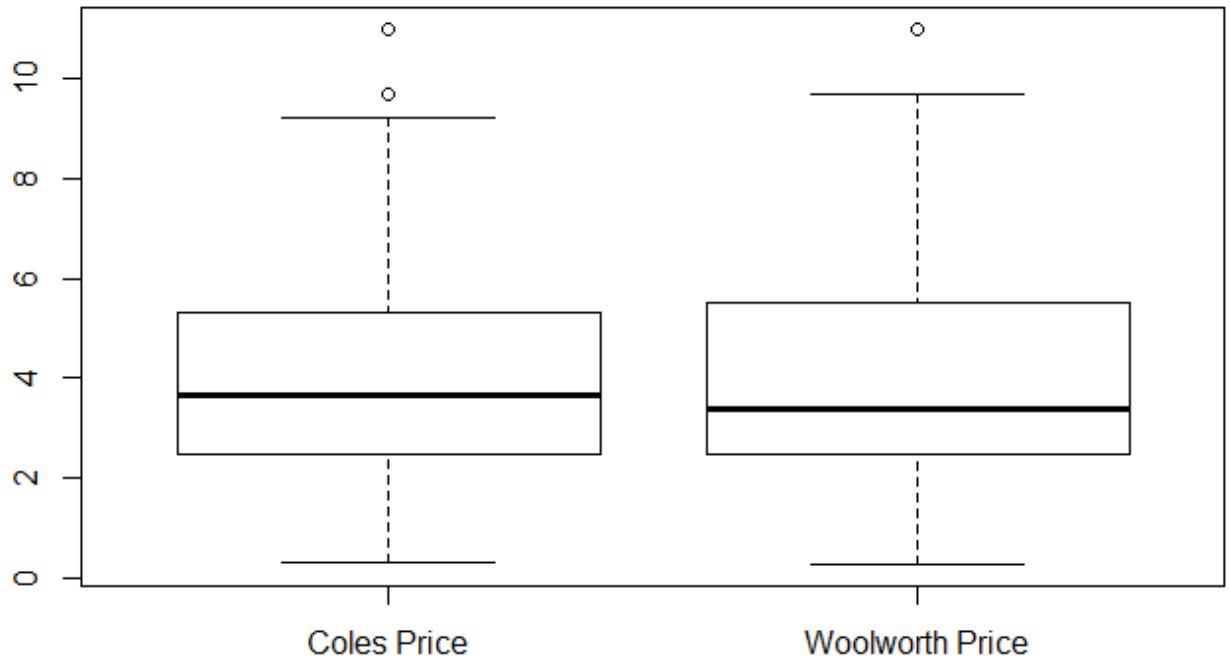
```
kable(WoolworthsPrices, caption = 'Summary Statistics for Woolworth Prices') %>% kable_styling
(bootstrap_options = c("hover","condensed"))
```

Summary Statistics for Woolworth Prices

Observations	Mean	Median	Standard Deviation	First Quantile	Third Quantile	Inter Quantile	Minimum	Maximum	Missing
180	4.002944	3.4	2.159945	2.5	5.5	3	0.29	11	0

Hide

```
boxplot(PriceWars[,3:4])
```



Hypothesis Test

To perform hypothesis test, we will use Levene's Test and two sample t-test. To perform Levene's test, a new variable called store is created to compare the prices based on grouping. The two sample t-test assumes the price samples collected from Coles and Woolworths are independent of each other. The default confidence level is considered for the test i.e. 95%. Significance level thus becomes 0.05 which is accepted as a standard in order to prove our Null hypothesis and we are considering μ=0 in each case. Thus, for Null Hypothesis, we assume both the Coles and Woolworths Supermarket have same prices.

Hide

```
#Q-Q plot for Coles
set.seed(1)
s <- sample_n(PriceWars,size = 180)
s
```

…	Products	Coles Price
<dbl⌅	<chr>	<dbl>
68	Wokka Noodles Golden Hokkien Shelf Fresh	2.70
167	Primo Chicken Breast Thinly Sliced	2.50

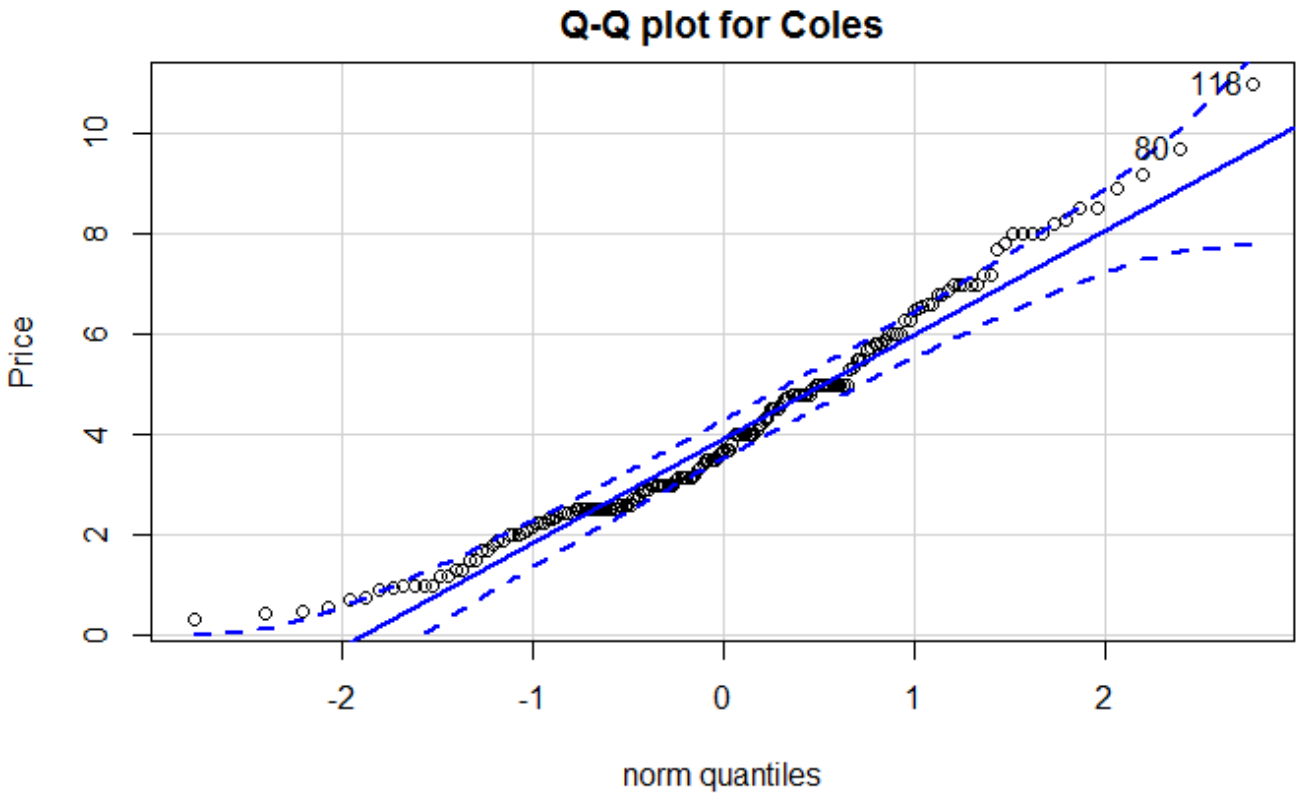
…	Products	Coles Price
	<dbl⧺chr>	<dbl>
129	Jalna Greek yoghurt	0.70
162	Fruche Vanilla Bean Dessert	2.99
43	Weight Watchers Frozen Creamy Tuna Bake Meal	5.80
14	Fresh Pet Food Co Pets Mince	0.55
51	Cadbury Baking White Chocolate Melts	4.19
85	Mount Franklin Lightly Sparkling Mineral Water	2.85
21	McCain Frozen Superfries Crinkle Cut Potato Chips	3.00
106	Original Juice Grapefruit Juice	5.00

1-10 of 180 rows | 1-4 of 5 columns Previous **1** 2 3 4 5 6 … 18 Next

Hide

```
s$`Coles Price` %>% qqPlot(dist="norm", main = "Q-Q plot for Coles", ylab = "Price")
```

```
[1] 118  80
```

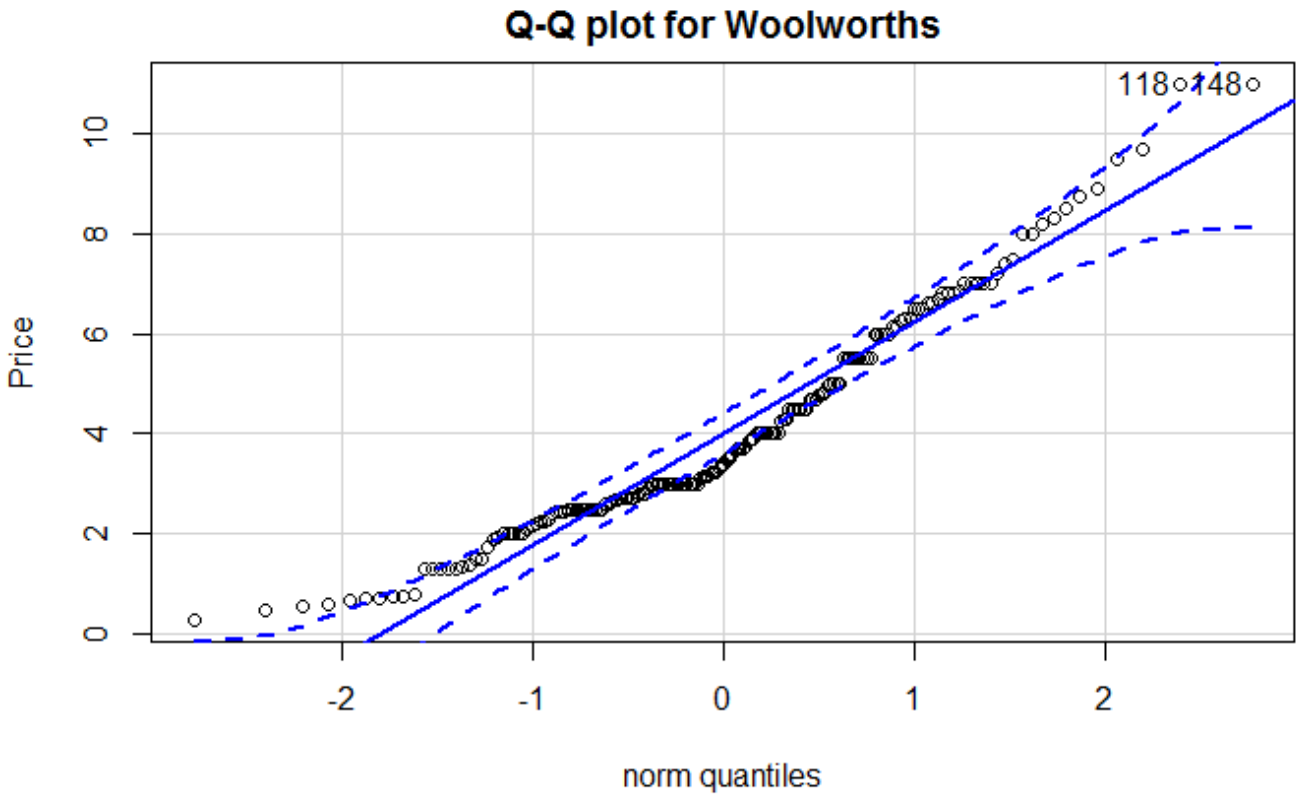


Hide

```
#Q-Q plot for Woolworths

s$`Woolworth Price` %>% qqPlot(dist="norm", main = "Q-Q plot for Woolworths", ylab = "Price")
```

```
[1] 118 148
```



Hide

```
#Levene's Test to check homogeneity of variance
Price <- c(PriceWars$`Coles Price`,PriceWars$`Woolworth Price`)

x1 <- rep('c',times = 180)
x2 <- rep('w',times = 180)
Store <- factor(x= c(x1,x2))

leveneTest(Price ~ Store, data = s)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.0265 0.8709
      358
```

Hide

```
#Indepentent two sample t-test for independent variables with equal variance
t.test(Price ~ Store,data = s,var.equal = TRUE,alternative = "two.sided")
```

```
    Two Sample t-test

data:  Price by Store
t = 0.23616, df = 358, p-value = 0.8134
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3907965  0.4974632
sample estimates:
mean in group c mean in group w
       4.056278        4.002944
```

Interpretation

Referring to the data collected and tests done, we can observe that there is not enough evidence to prove that either one of the supermarket is costly or cheaper. For few category of products Coles is expensive and for the rest Woolworths is expensive. Thus, we cannot completely determine which supermarket is cheaper among Coles and Woolworths by comparing results from different categories.

Two sample t-test was used to determine the significant difference between the mean prices of Coles and Woolworths. The results of the two-sample t-test assuming equal variance found the difference between the mean prices of Coles and Woolworths are not statistically significant, t(df=358)= 0.23616, p=0.81, 95% CI for the difference in means [-0.39 0.50].The results of the investigation suggest that to find which of the two supermarkets, Woolworths or Coles is cheaper, requires more evidence to deduce the statistically significant results.

Discussion

As per our analysis, we came to the conclusion that there is no statistically significant difference in the mean of both the supermarkets to suggest which is cheaper than the other. However our analysis was limited to the random samples we collected.

The strengths of this analysis is that we choose the most common categories of the items that are bought by most of the buyers by keeping an upper limit of approximately 11$. As far as the weaknesses are concerned, there were a fair number of outliers visible in the Q-Q plots of both supermarkets. Also, we feel that the sample size could have been bigger.

Finally, we believe that the larger dataset will result in more clear and transparent analysis and the investigation will turn out to be more reliable. Statistically, more the data, more significant will be the analysis.