MATH1324 Assignment 2

Supermarket Price Wars

Individual Details

- Vinay Nagamanagala Ramegowda

Executive Statement

- The aim of this investigation is to compare the prices of two of the well-known supermarkets and draw conclusions as to which supermarket is cheaper than the other. The data collected was compiled directly from their website and manually from their stores. Sixty-five different products were randomly observed in total and we made sure to select only those products which are available in both the stores. The prepared data is then imported into Rstudio.

- To describe the data, we obtained summary statistics of the entire data and for better visualization we plotted a boxplot for each supermarket. A paired t-test is run on the dataset to figure which of the two retailers have the cheaper prices. To run paired t-test we must ensure the data collected has equal variance and is normal. Since our data set has more than 30 records it is considered as normal according to Central Limit Theorem. Then a Levene test with a significance level of 0.05 is run to check the homogeneity of variances among the two retailers. Firstly, we perform Fisher test to check whether there is any homogeneity in the data.

- Later on, if we perform t-test ,the p-value of the dataset is analysed with the significance level of 0.05. After performing the paired t-test on the data we get p-value as 0.01665 which is less than the significance value of 0.05 and hence is statistically significant.

Load Packages and Data

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(lawstat)
library(PairedData)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: gld
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'PairedData'
```

```
## The following object is masked from 'package:base':
##
##     summary
```

```
supermarket_prices<-read.csv(file.choose(),header = TRUE)
supermarket_prices
```

ï..product_name <fctr>	woolies_pric <dbl
4 Pines Brewing Company Pale Ale Bottles 6x330ml pack	24.0
5 Seeds Apple Cider Crisp Cans 10x330ml pack	22.0
5 Seeds Lower Sugar Cider 24x345ml case	25.0
Brewmanity Social Beast Pale Ale Can 375ml x6 pack	24.5
Bridge Road Brewers Beechworth Pale Ale Bottles 6x330ml pack	25.0
Carlton Dry Bottle 330ml x24 pack	28.0

ï..product_name <fctr>	woolies_pric <dbl
Carlton Dry Lime Bottle 24x330ml	28.0
Carlton Zero Bottle 330ml x24 pack	20.0
Coopers Mild Ale Stubbies 24x375ml case	25.0
Coopers Pale Ale Can 4x6x375ml	25.0

1-10 of 65 rows Previous **1** 2 3 4 5 6 7 Next

Summary Statistics

We first find summary statistics of each of the supermarkets by using the summarise function. Then we plot a box-plot for better visualization of the data.

```
supermarket_prices %>% summarise(LowestPrice = min(`woolies_price`),
                         HighestPrice = max(`woolies_price`),
                         Q1 = quantile(`woolies_price`, probs=0.25),
                         Median = median(`woolies_price`),
                         Q3 = quantile(`woolies_price`, probs=0.75),
                         Mean = mean(`woolies_price`),
                         StandardDeviation = sd(`woolies_price`),
                         TotalRecords = n())
```

LowestPrice <dbl>	HighestPrice <dbl>	… <dbl>	Med… <dbl>	Q3 <dbl>	Mean <dbl>	StandardDeviation <dbl>	TotalRecords <int>
5	30	7.7	15	24.5	15.76369	8.270696	65

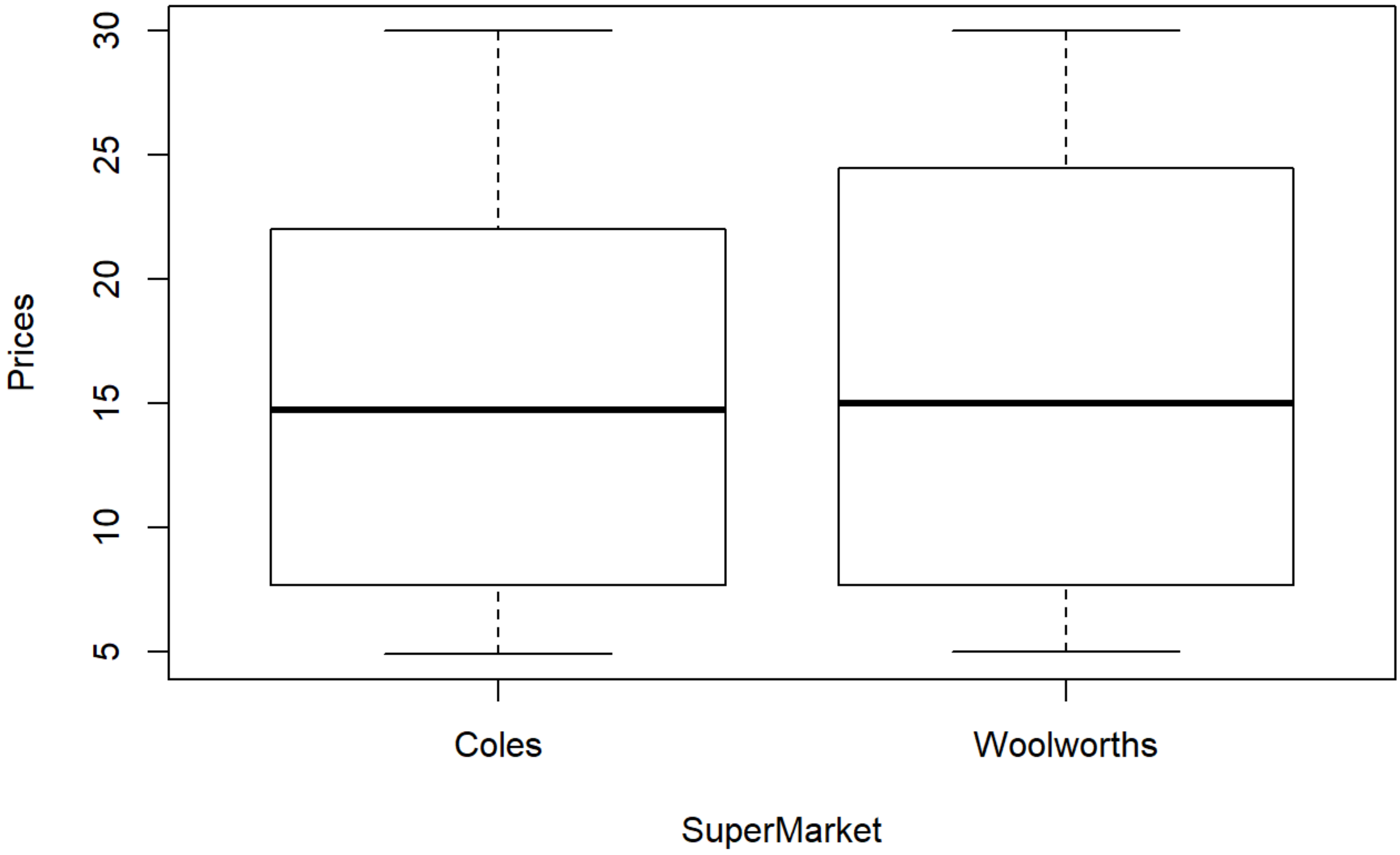
1 row

```
supermarket_prices %>% summarise(LowestPrice = min(`coles_price`),
                         HighestPrice = max(`coles_price`),
                         Q1 = quantile(`coles_price`, probs=0.25),
                         Median = median(`coles_price`),
                         Q3 = quantile(`coles_price`, probs=0.75),
                         Mean = mean(`coles_price`),
                         StandardDeviation = sd(`coles_price`),
                         TotalRecords = n())
```

LowestPrice <dbl>	HighestPrice <dbl>	… <dbl>	Med… <dbl>	… <dbl>	Mean <dbl>	StandardDeviation <dbl>	TotalRecords <int>
4.9	30	7.7	14.75	22	14.85723	7.553785	65

1 row

```
boxplot(supermarket_prices$coles_price,
        supermarket_prices$woolies_price,
        ylab = "Prices",
        xlab = "SuperMarket", names = c("Coles","Woolworths"))
```



Hypothesis Test

```
var.test(supermarket_prices$woolies_price, supermarket_prices$coles_price, alternative = "two.sided") #Fisher f-test to check the homogeneity of the variances
```

```
## 
##  F test to compare two variances
## 
## data:  supermarket_prices$woolies_price and supermarket_prices$coles_price
## F = 1.1988, num df = 64, denom df = 64, p-value = 0.4703
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.731219 1.965452
## sample estimates:
## ratio of variances 
##           1.198823
```

```
t.test(supermarket_prices$woolies_price,supermarket_prices$coles_price,alternative = "two.s
ided", paired=T) #t-test check for statistical significance
```

```
##
##  Paired t-test
##
## data:  supermarket_prices$woolies_price and supermarket_prices$coles_price
## t = 2.4589, df = 64, p-value = 0.01665
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1700014 1.6429217
## sample estimates:
## mean of the differences
##               0.9064615
```

Interpretation

- To figure out which one of the supermarkets is cheaper we investigate the data on two-tailed paired-sample t-test. In the result of Fisher test we get p-value as 0.4703 which is much greater than the significance level of 0.05. Hence the homogeneity of variances holds true and we can now perform t-test.
- After performing t-test we get p-value 0.01665 which is lesser than significance level 0.05, hence we reject null hypothesis and conclude that there is significant difference between the prices of the two retailers.

Discussion

The data collected is just a mere fraction of the entire data and also the data gathered is based on our preferences. Since data collection is not efficient and not large enough , we cannot come to any conclusion. we will need a much larger data set comprising of a wide variety of products to come to a better conclusion.