Code -

Supermarket Price Wars

Mujeer M.

Coles v. Woolworths

Executive Statement

This report aims to provide a statistically apt comparison between the prices of Coles and Woolworths. The intended results in conducting this exercise were to determine how Coles and Woolworths differ across various price points within a constrained set of products which were randomly sampled. These products range across multiple categories and help give us a holistic reflection of our data set.

In order to bring our exercise to fruition a chronological set of tasks were conducted. These included the following in order:

- Collection of data.
- Categorical sampling of data.
- Population statistic analysis across price points between the supermarkets.
- Hypothesis testing - to determine statistical significance of our assumptions on the data.
- Interpretation of results.
- Discussion of results.

The data collection was done by visiting the websites of both the supermarkets 21/09/2018 to ensure live data was procured. For conclusive results we also ensured that the data for products collected were same across Woolworths and Coles in terms of make, model and unit size.

The data collection was done under 5 categories i.e. Liquor, Health and Beauty, Fruits and Veggies, Pantry and Dairy. Random sampling was done under each category and these samples were stitched together to create a final data set. Due to this collection and sampling methodology that was implemented, we believe that our final data set:

- Accounts for the necessary randomisation.
- Is a representation of the population.

Post the collection stage we needed to visualise the data to have a preliminary understanding of what our analysis was pointing towards. However, as the word reflects, this was a preliminary outcome and did not provide conclusive evidence. To achieve that clarity, we conducted paired t-Test. We set the Null hypothesis as the mean difference between paired observations is zero. The results of the test are shown later in the report.

Load Packages and Data

Hide

```
library("readr")
library("xlsx")
library("magrittr")
library("dplyr")
library("ggplot2")
library("car")
```

Hide

```
mas_health <- read.xlsx("Coles & WW.xlsx","Health and Beauty") %>% as.data.frame() %>% na.omi
t()
mas_fruits <- read.xlsx("Coles & WW.xlsx","Fruits and Veggies") %>% as.data.frame() %>% na.om
it()
mas_pantry <- read.xlsx("Coles & WW.xlsx","Pantry") %>% as.data.frame() %>% na.omit()
mas_dairy <- read.xlsx("Coles & WW.xlsx","Dairy") %>% as.data.frame() %>% na.omit()
```

Hide

```
sam_health <- sample_n(mas_health,10)
sam_fruits <- sample_n(mas_fruits,10)
sam_pantry<- sample_n(mas_pantry,10)
sam_dairy <- sample_n(mas_dairy,10)
```

As discussed in the Executive Summary section of this report, sample_n function was used to randomly select 10 products from each category.

Hide

```
final_data <- rbind(sam_liq,sam_health,sam_fruits,sam_pantry,sam_dairy)
head(final_data)
```

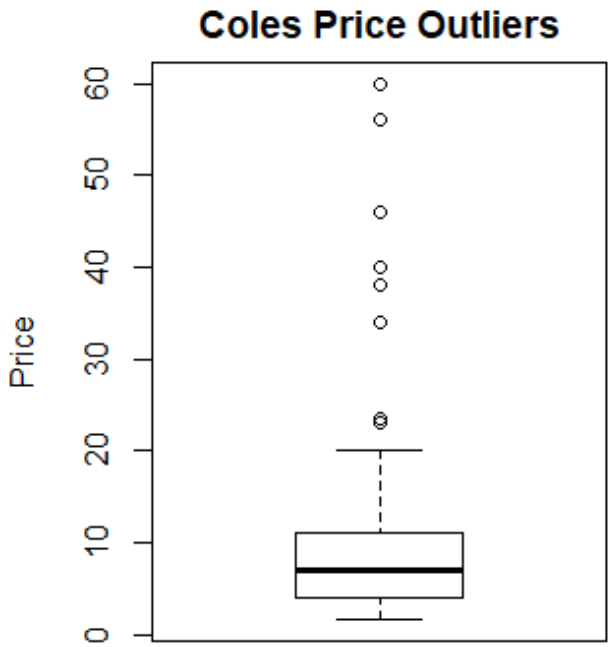
	Category <fctr>	Product.name <fctr>	Coles.Price <dbl>	Woolworth.s.Price <dbl>
18	Liquor	Yellow tail sparkling bubbles 750ml	10.0	10.0
20	Liquor	Jagermeister 700ml	56.0	40.0
14	Liquor	Southern comfort bourbon 700ml	38.0	38.0
1	Liquor	Corona extra 6 pack 355ml	23.5	22.5
15	Liquor	Canadian club whiskey 700ml	38.0	38.0
11	Liquor	Bundaberg Rum 200ml	23.0	22.0

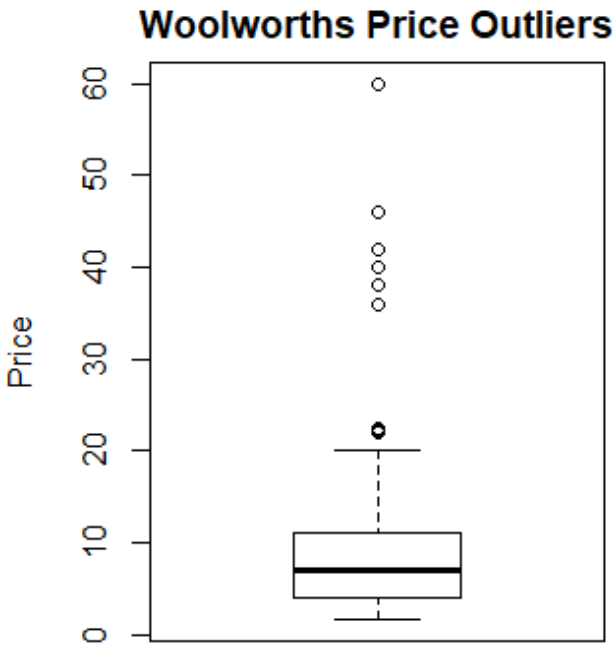
6 rows

rbind function was used to stitch these categorically sampled data sets into one data frame i.e. final_data.

Hide

```
final_data$Coles.Price %>% boxplot(main="Coles Price Outliers",ylab="Price")
final_data$Woolworth.s.Price %>% boxplot(main="Woolworths Price Outliers",ylab="Price")
```

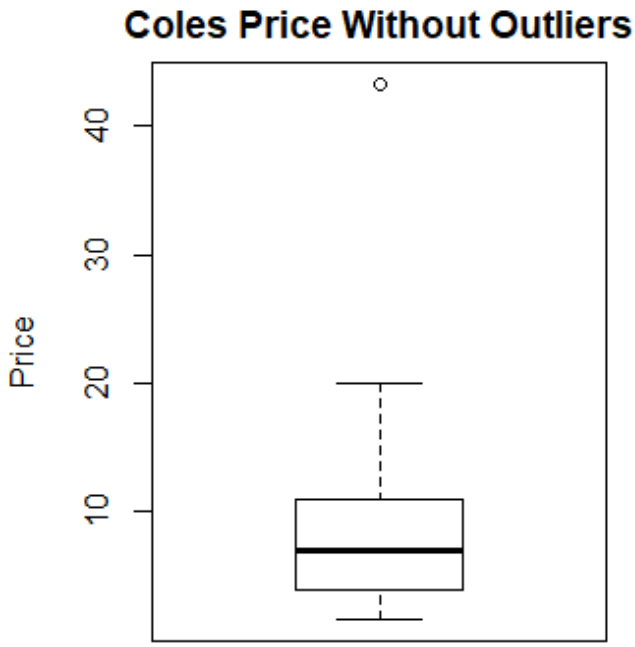


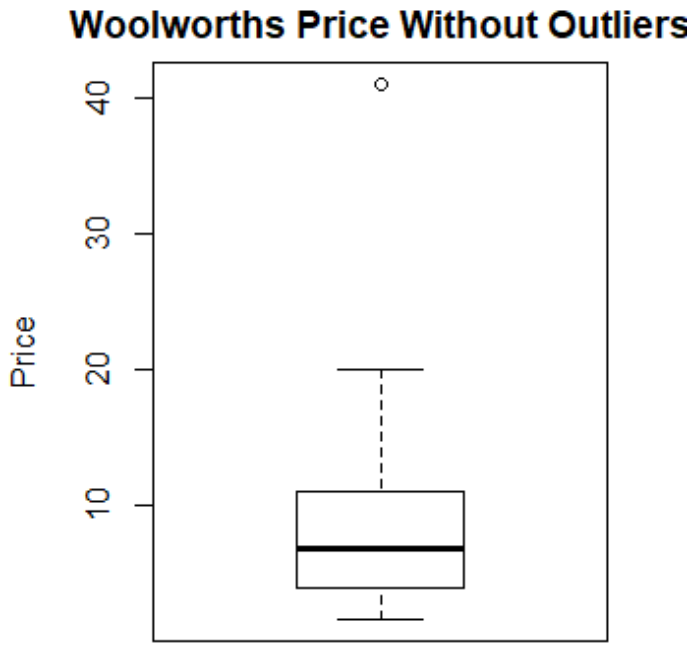


We can see that the price varianbles of Coles and Woolworths have outliers.

Hide

```
cap <- function(x){
  quantiles <- quantile( x, c(.05, 0.25, 0.75, .95 ))
  x[ x < quantiles[2] - 1.5*IQR(x) ] <- quantiles[1]
  x[ x > quantiles[3] + 1.5*IQR(x) ] <- quantiles[4]
  x
}
final_data$Coles.Price <- final_data$Coles.Price %>% cap()
final_data$Woolworth.s.Price <- final_data$Woolworth.s.Price %>% cap()
par(mfcol=c(1,2))
final_data$Coles.Price %>% boxplot(main="Coles Price Without Outliers",ylab="Price")
final_data$Woolworth.s.Price %>% boxplot(main="Woolworths Price Without Outliers",ylab="Pric
e")
```





We used the Windsorising/Capping methodology to deal with these outliers.

Hide

```
final_data <- final_data %>% mutate(Price_Difference=(final_data$Coles.Price - final_data$Woo
lworth.s.Price))
head(final_data)
```

	Category <fctr>	Product.name <fctr>	Coles.Price <dbl>	Woolworth.s.Price <dbl>	Pri
1	Liquor	Absolut 700ml	54.875	50.95	
2	Liquor	Bacardi Rum 1L	54.875	50.95	
3	Liquor	Sierra Tequila 700ml	54.875	50.95	
4	Liquor	Corona extra 6 pack 355ml	54.875	22.50	
5	Liquor	Coopers pale ale 6 pack 375ml	19.000	20.00	
6	Liquor	Peroni Nastro Azzuro 6 pack 330ml	54.875	22.30	

6 rows

We created a new column called Price_Difference witin the final_data data frame using the mutate function. This column will contain the difference of prices for each product across both the supermarkets.(We did Coles Price - Woolworths Price).

Summary Statistics

Use R to summarise the data from the investigation. Include an appropriate plot to help visualise the data. Describe the trend.

Hide

```
lapply(final_data[,3:5], summary)
```

```
$`Coles.Price`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.840   3.848  10.863  54.875                  

$Woolworth.s.Price
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.500   4.143   7.245  13.011  13.750  50.950 

$Price_Difference
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -4.500  -0.485   0.000   3.051   1.845  39.905 
```

Summary stats(except sd, IQR) have been derived for Price across Coles,Woolworths in addition to the Price Difference column.

```
$`Coles.Price`
[1] 19.8854

$Woolworth.s.Price
[1] 15.09632

$Price_Difference
[1] 9.561051
```

Standard Deviation has been derived for Price across Coles,Woolworths in addition to the Price Difference column.

Hide

```
lapply(final_data[,3:5],IQR)
```

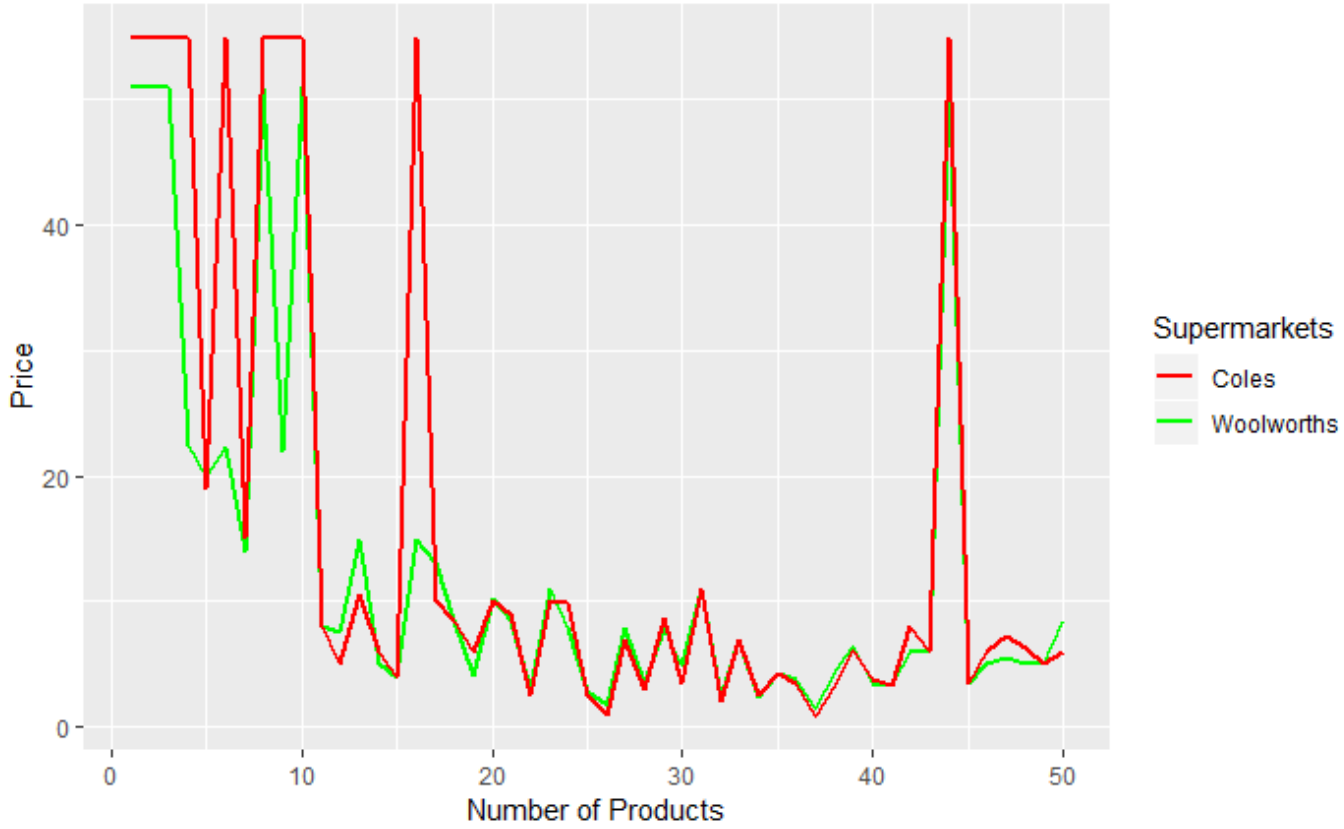
```
$`Coles.Price`
[1] 7.015
 9.6075

$Price_Difference
[1] 2.33
```

IQR has been derived for Price across Coles,Woolworths in addition to the Price Difference column.

Hide

```
  geom_line(data=final_data,aes(x=1:50,final_data$Coles.Price,col="Coles"),size=1) +
  scale_color_manual(name = "Supermarkets",
                     values = c("Woolworths" = "green", "Coles" = "red"))+labs(x="Number of P
roducts",y="Price")
```



Hide

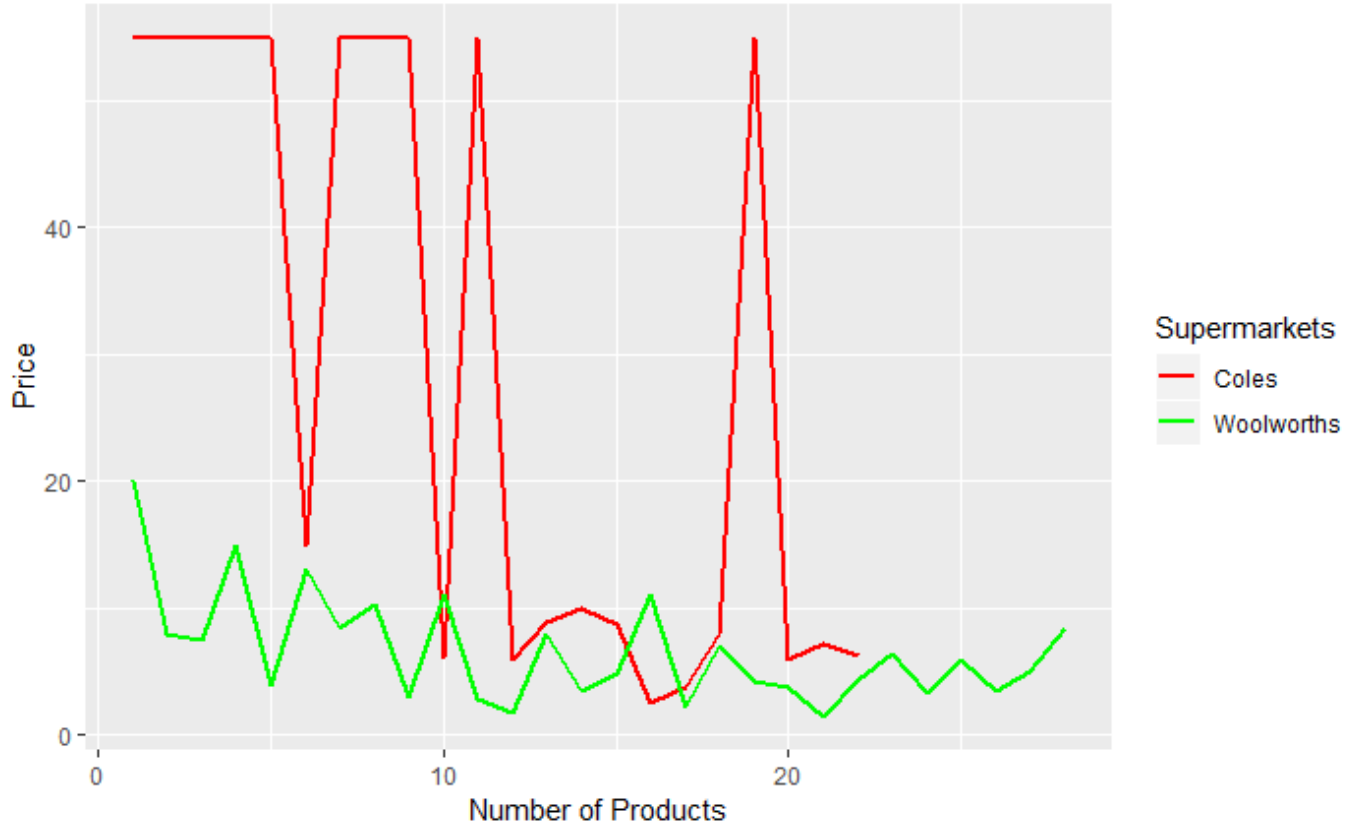
```
coles_high_cost <- final_data %>% filter(Price_Difference > 0)
wool_high_cost <- final_data %>% filter(Price_Difference <= 0)
```

The above visualization plots the prices of products across coles and woolworths. Here, we can observe that coles has a slight gain with respect to prices over woolworths.

The next visualization sheds more ligh on this insight.

Hide

```
ggplot() + geom_line(data=coles_high_cost , aes(x=1:22,coles_high_cost$Coles.Price,col="Cole
s"),size=1)+
  geom_line(data=wool_high_cost , aes(x=1:28,wool_high_cost$Woolworth.s.Price,col="Woolworth
s"),size=1) +
   scale_color_manual(name = "Supermarkets",
```



Looking at the above visualization we can see that coles has a higher price point in comparision to woolworths, amongst the products with higher price across coles and woolworths.

This indicates that Coles is costly in comparision with Woolworths. We will be conducting a paired t-Test to understand the validity of this claim.

Hypothesis Test

We are using the Paired T test to conduct this Hypothesis. This is because:

- This price related data has been collected for the same product across 2 supermarkets i.e the subjects(products) in the first group are also in the second group
- The data is numeric and continuous.
- No outliers in the Price_Difference column.

Paired Sample t-Test:

Null Hypothesis: The mean difference between paired observations is zero

Hypothesis: The mean difference between paired observations is not equal to zero.

Normality Assumption:

The Price_Difference is normally distributed as the sample size is 50 (i.e > 30). This has been assumed in referrence to Central Limit Theorem.

Hide

```
t.test(final_data$Coles.Price, final_data$Woolworth.s.Price,
       paired = TRUE,
       alternative = "two.sided")
```

```
    Paired t-test

data:  final_data$Coles.Price and final_data$Woolworth.s.Price
t = 2.2567, df = 49, p-value = 0.02852
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3341795 5.7686205
sample estimates:
mean of the differences
                 3.0514
```

Looking at the results of the above, we reject the null hypothesis as $p < 0.05$. This means that the results of this Hypothesis Test are statistically significant.

Interpretation

A paired-samples t-test was used to test for a significant mean difference between products prices across Coles and Woolworths. The mean difference following exercise was found to be 3.0514 (SD= 9.56). The paired-samples t-test found a statistically significant mean difference between product prices across both the supermarkets, t(df=49)=2.2567, p = 0.02852 , 95% [0.3341795 5.7686205].

From our Hypothesis Testing we can conclude that Woolworths is cheaper than Coles by a mean Price Difference of 3.0514 AUD.

Discussion

Before our findings can be discussed in their entirety a few variables need to be considered for appropriate understanding of said findings. These will be presented below as assumptions and limitations.

Another important note to add here would be our reasoning behind choosing the paired t-test instead of the unpaired t-test. A very simple notion which was able to guide us in making this decision was that the prices which were collected despite belonging to different supermarkets represented the same makes , unit sizes and brands thus making them dependent.

Assumptions:

The tests and hypothesis were conducted based on the premise that the following statements held true for our data sample:

- On the date of data collection i.e 22/09/2018 websites for Coles and Woolworths presented accurate and live prices for their products.
- On the date of data collection Coles and Woolworths were not running any specials.
- The prices collected were uniform across Australia and did not vary across stores, suburbs and states.
- Our data set was a true representation of the population.

Limitations:

Because our data set is representative of convenience sampling it limits the notion of our assumption that our data set is representative of the entire population.

Furthermore a few other aspects need to be considered in viewing our results:

- It is likely that on the day of data collection prices reflected on the relevant websites had not been updated.
- That the prices across suburbs and states in fact do vary and that the prices are not uniform across Australia.

- Finally, there were special prices for various products on both the websites.

These limitations may have had a restrictive effect on our findings however as they are not able to be mitigated we have considered our assumptions to hold true.

Before we conclude this analysis, please find below some of the aspects that could be considered to improve the insights, for the next run:

- Adding an extra variable to specify whether a product is in special price and inculcating this aspect into our analysis.
- Collection of data over a period of time and including this date/time variable in our dataset could give us an extra dimension for the analysis.

Results :

Ultimately based on the assumptions taken into consideration and the limitations we faced during the conduct of this investigation we can conclude that price of Woolworths products are statistically significantly cheaper than that of Coles products by a mean price difference of 3.0514 AUD.