MATH1324 Assignment 2

Supermarket Price Wars

Group Details

- Ancy Rex (S3760058)
- Pooja Suresh (S3749775)
- Sridevi Pamarthi (S3778317)

Executive Statement

Aim Goal of investigation is to compare the price(AUD) of various products in two supermarkets Coles and Woolworths to find which is cheaper.

About the Sample

- To acheive our goal 225 items were choosen from both Coles and Woolsworths at random through their official website.
- To ensure the price difference(AUD) between the items the same brand and category were compared and choosen to include into the dataset.
- A dataset of 225 items were taken so as to obtain a normal distribution.
- The price of items on sale and offers were ignored to obtain an unbiased dataset.

Procedure Followed

- The dataset is dependent since it consists of same products from the two supermarkets.
- The null hypothesis H0 was considered as mean difference in the price is 0 (HO=mu1-mu2) and alternate hypothesis HA as mean difference in the price is not equal to 0.
- Boxplots and qqplot were used to find the normality in prices
- To support our hypothesis we have used paired sample t-test to determine value the value of p.
- If the p value is less than alpha, we reject null hypothesis otherwise, we fail to reject the null hypothesis.

Load Packages and Data

```
#Required Packages

library(dplyr)
library(readr)
library(magrittr)
library(car)

# Importing data

Price.war.Paired <- read_csv("Price - Coles vs. Woolworths.csv")
View(Price.war.Paired)

# Create differences (d) column

Price.war.Paired <- Price.war.Paired %>% mutate(d =Coles-Woolworths)
```

Summary Statistics

The Summary Statistics was done to compare the prices of the products in both supermarkets (Coles and Woolworths).

- Boxplots were produced to understand the normality in the prices.
- It was brought to attention that the prices did fall under the 95% CI.
- According to Central Limit Theorem(CLT), if sample size is greater than 30 it means data is approximately normal. As sample size is 225, according to CLT, data is normal.

```
# Summary Statistics for Coles
Price.war.Paired %>% summarise(Mean = mean(Coles, na.rm = TRUE),
                               Median = median(Coles, na.rm = TRUE),
                               SD = sd(Coles, na.rm = TRUE),
                               Q1 = quantile(Coles, probs = .25, na.rm = TRUE),
                               Q3 = quantile(Coles, probs = .75, na.rm = TRUE),
                               Min = min(Coles, na.rm = TRUE),
                               Max = max(Coles, na.rm = TRUE),n = n())
```

```
## # A tibble: 1 x 8
##    Mean Median    SD    Q1    Q3   Min   Max     n
##   <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>
## 1  6.82      5  5.43     3     9  0.95    40   225
```

```
# Summary Statistics for Woolworths
Price.war.Paired %>% summarise(Mean = mean(Woolworths, na.rm = TRUE),
                               Median = median(Woolworths, na.rm = TRUE),
                               SD = sd(Woolworths, na.rm = TRUE),
                               Q1 = quantile(Woolworths,probs=.25,na.rm=TRUE),
                               Q3 = quantile(Woolworths,probs=.75,na.rm =TRUE),
                               Min = min(Woolworths, na.rm = TRUE),
                               Max = max(Woolworths, na.rm = TRUE),n = n())
```

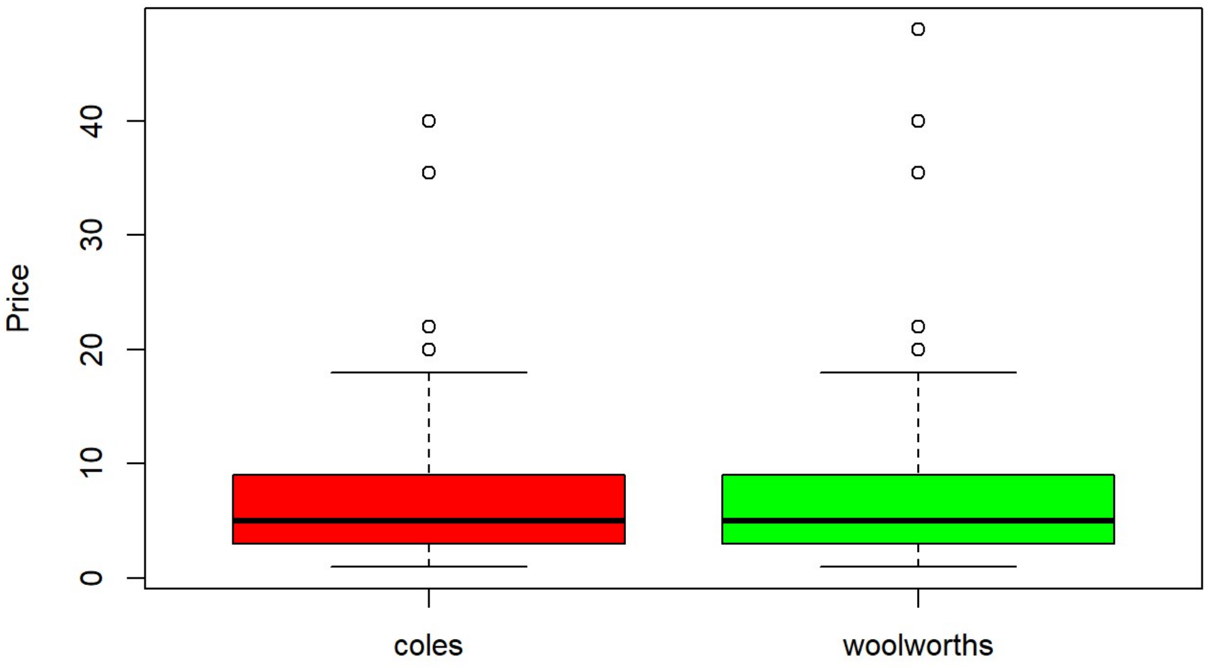
```
## # A tibble: 1 x 8
##    Mean Median    SD    Q1    Q3   Min   Max     n
##   <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>
## 1  6.85      5  6.06     3     9  0.95    48   225
```

```
Price.war.Paired %>% summarise(Mean_Coles = mean(Coles, na.rm = TRUE),
                     SD_Coles = sd(Coles, na.rm = TRUE),
                     Mean_Woolworths = mean(Woolworths, na.rm = TRUE),
                     SD_Woolworths = sd(Woolworths, na.rm = TRUE),
                     Mean_Difference = Mean_Coles - Mean_Woolworths,
                     SD_Difference = sd(Coles-Woolworths, na.rm = TRUE),
                     n = n())
```

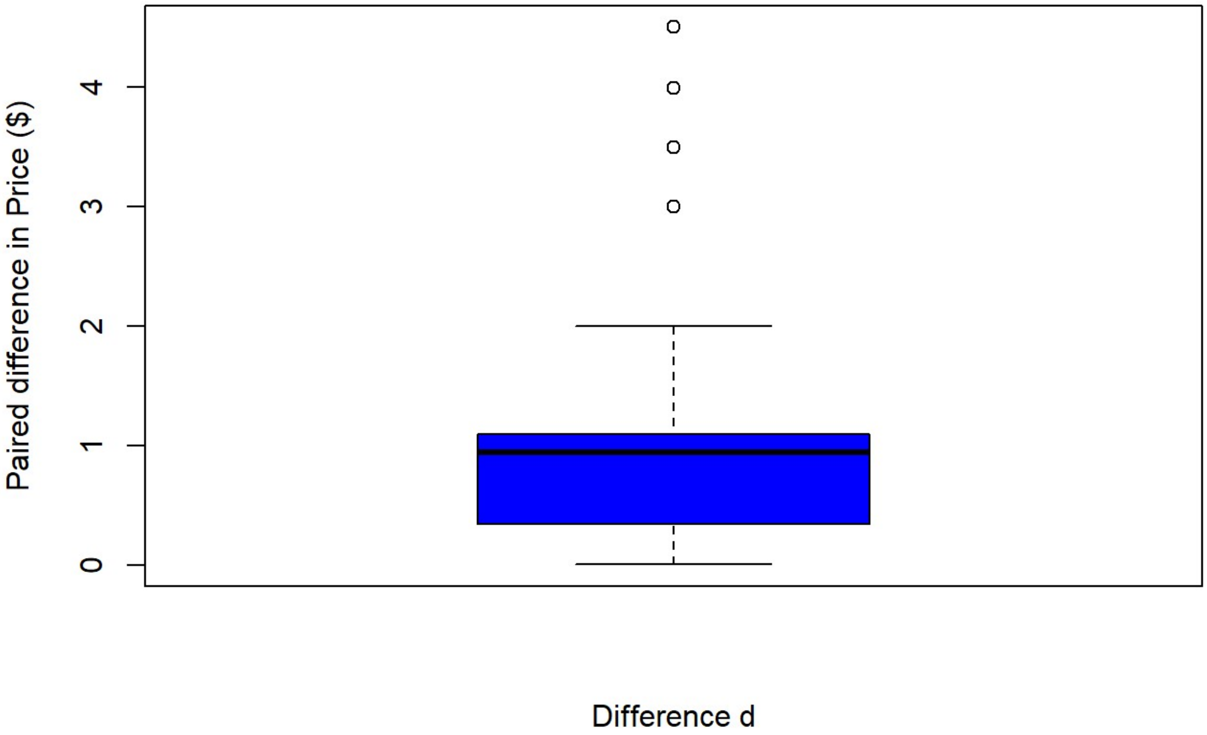
```
## # A tibble: 1 x 7
##   Mean_Coles SD_Coles Mean_Woolworths SD_Woolworths Mean_Difference
##        <dbl>    <dbl>           <dbl>         <dbl>           <dbl>
## 1       6.82     5.43            6.85          6.06         -0.0318
## # ... with 2 more variables: SD_Difference <dbl>, n <int>
```

```
# Visualisation

boxplot(Price.war.Paired$Coles,Price.war.Paired$Woolworths,ylab = "Price",col=c('re
d','green'))
axis(1, at = 1:2,labels = c("coles", "woolworths"))
```

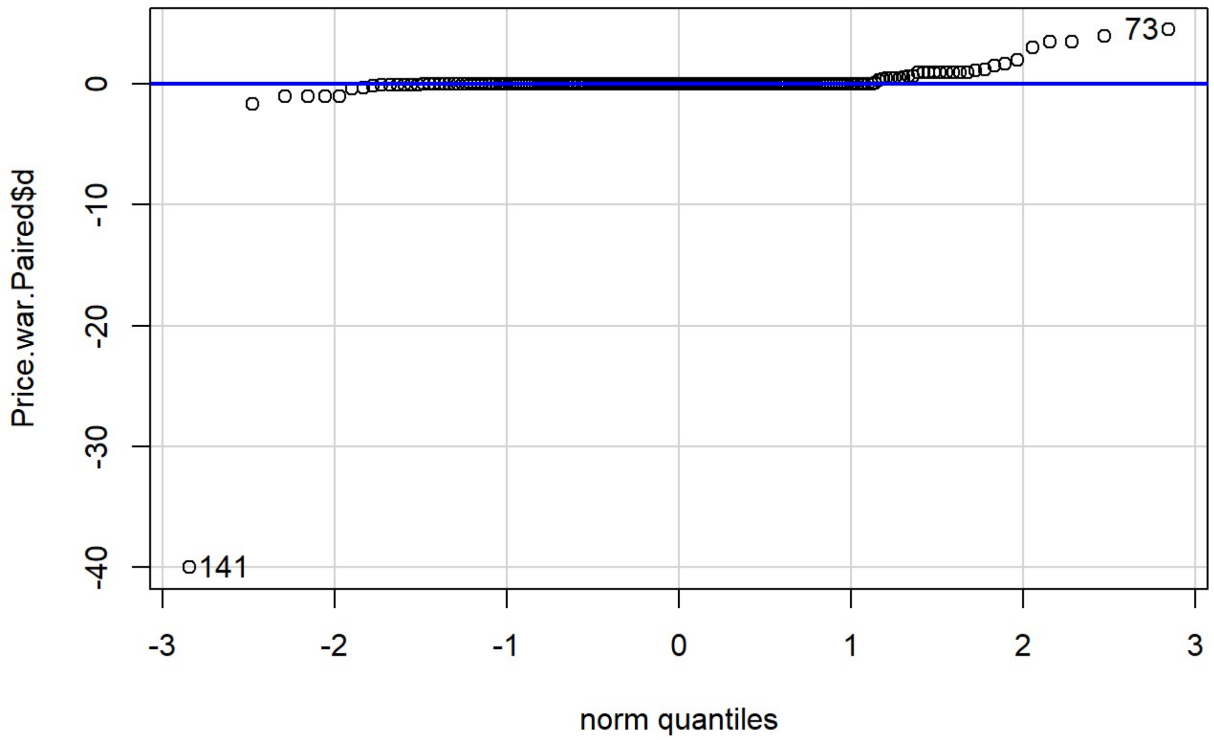


```
#Filtering the outliers
Price.war.Paired_Filter <- Price.war.Paired %>% filter(d >0)
boxplot(Price.war.Paired_Filter$d,col='Blue',ylab="Paired difference in Price ($)", xl
ab= "Difference d")
```



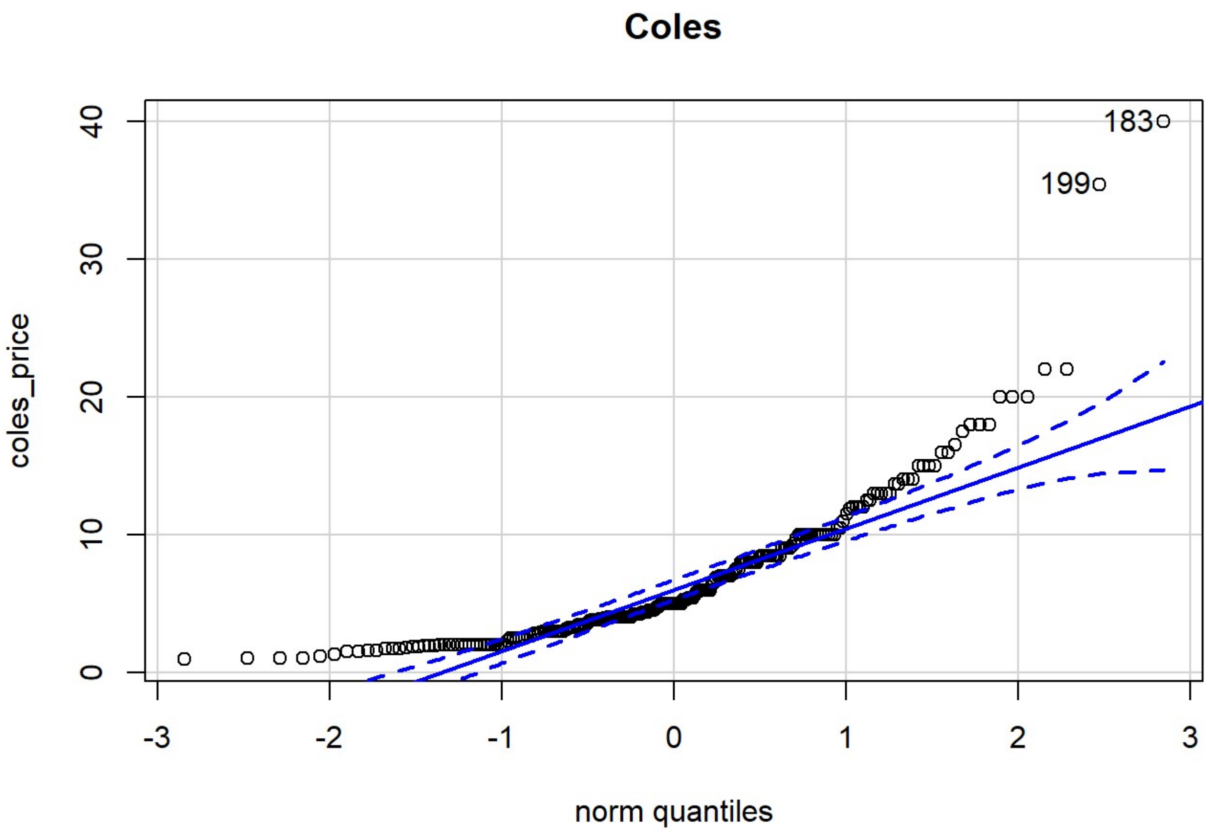
```
#Q-Q plot

library(car)
qqPlot(Price.war.Paired$d,dist="norm")
```



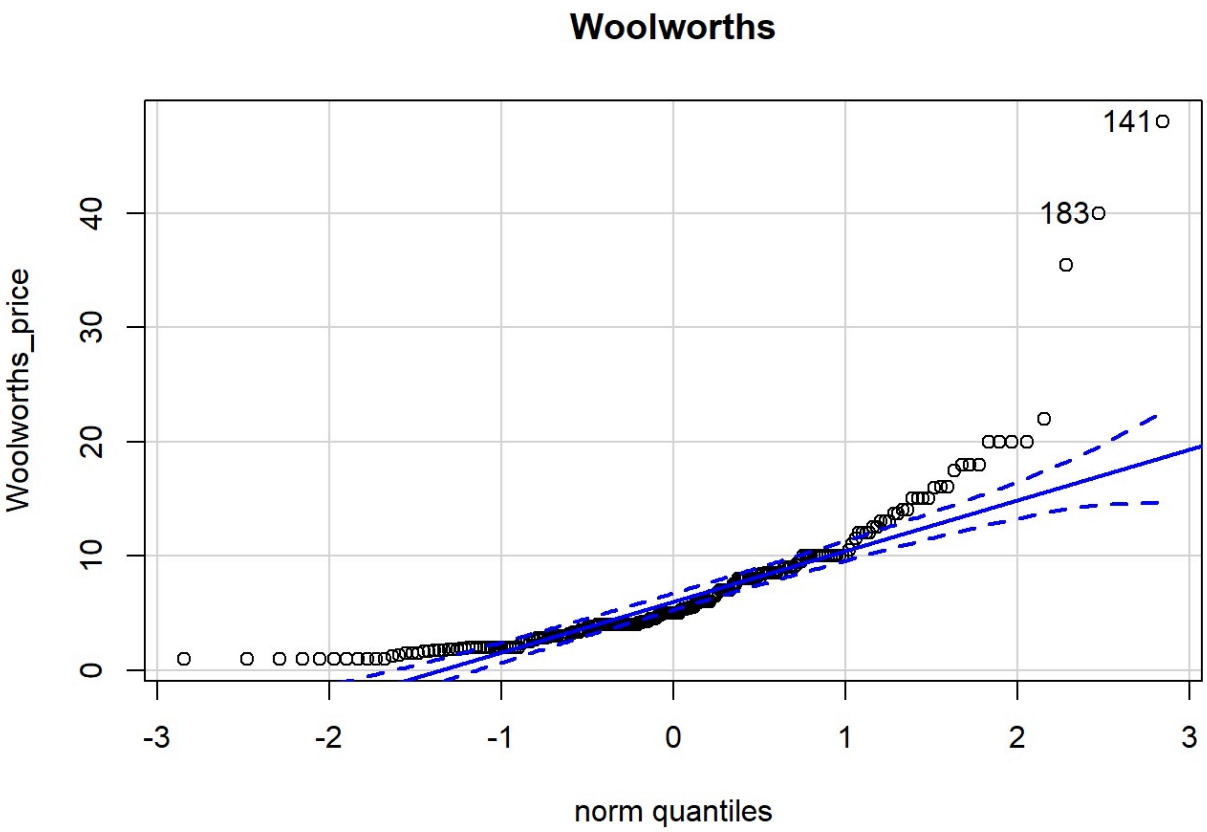
```
## [1] 141  73
```

```
#Q-Q plot Coles
qqPlot(Price.war.Paired$Coles,dist="norm",main="Coles",ylab="coles_price")
```



```
## [1] 183 199
```

```
#Q-Q plot Woolworths
qqPlot(Price.war.Paired$Woolworths,
        dist="norm",main="Woolworths",ylab="Woolworths_price")
```



```
## [1] 141 183
```

Hypothesis Test

- The data is dependent because they are compared among the same products.
- The paired sample t-test was done to obtain the p value so as to reach the hypotheis conclusion.

Null Hypothesis The mean value of both the supermarkets are equal.(Ho = mu1-mu2 = 0)

Alternate Hypothesis The mean value of both the supermarkets are not equal.

```
# hypothesis testing code.

# Conduct Paired sample t-test
pttest <- t.test(Price.war.Paired$Coles, Price.war.Paired$Woolworths,
                 paired = TRUE,alternative = "two.sided",conf.level = .95)

pttest
```

```
##
##  Paired t-test
##
## data:  Price.war.Paired$Coles and Price.war.Paired$Woolworths
## t = -0.17329, df = 224, p-value = 0.8626
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3937065  0.3300620
## sample estimates:
## mean of the differences
##             -0.03182222
```

Interpretation

The findings from results of the hypothesis test by acquiring the p-value and confidence intervals.

- The paired sample t-test was done for analysis.
- The t-test resulted in a p-value of 0.8626 which is greater than the alpha value 0.05.
- The mean difference falls in the 95% of CI.
- The result is not statistically significant.

Hence we come to the conclusion that:

- A paired-sample t-test was used to test for a significant mean difference between prices in Coles and Woolworths. The mean difference of the dataset was found to be -0.031822(SD = 2.75461).
- Visual inspection of the Q-Q plot of the difference scores suggested that the data was approximately normally distributed.
- The paired-samples t-test found a not statistically significant mean difference between prices,t (df=224) is -0.17329,p-value is 0.8626(>0.05) and 95% confidence interval[-7.08 -0.79].
- Hence, we Fail to reject the null hypothesis.
- There is not enough evidence to support our alternative hypothesis and thus the result is not statistically significant.
- Hence we conclude by saying that there is no statistical significance to indicate which supermarket is cheaper.

Discussion

The analysis has the following findings :

- The mean of the prices was done and found to be similar.
- There is not much difference in the product prices in both Coles and Woolworths.

Strengths and Limitations to our investigation:

- We have randomly collected a large sample to conduct the investigation.
- The dataset taken by choosing different products from both Coles and Woolworths helps study the comparison of both the supermarkets. The products are compared and found to have similar variety and price ranges.

- With larger amounts of data, this experiment would have lesser error ratio, greater efficiency and strucutre.
- Prices may vary from time to time. If we conduct hypothesis testing for different time periods, the results will be more efficient.
- Also comparing products by particular categories helps for more intresting findings.