# MATH1324 Assignment 2

Code

## Coles and Woolworths Supermarket Price Wars

# Group/Individual Details


- Ayesha Hojage(s3802861)
- Komal Mehta(s3795392)


# Executive Statement

Experiment Objective:-

The experiment compares the price rate of two of the Australia's well-known supermarkets 'Coles' and 'Woolworths' and finds out which supermarket provides their products in cheap price.

Sample Size:-

- Here we have used random sampling while collecting the products for our investigation.
- A random number was generated and a product was seleted for the sample based on this number.
- A sample of 49 Observations are collected to form the dataset, so as to cover the larger population mean.
- Due to large dataset,normally distributed curve of the complete datasetis formed.

Variables Taken in the dataset are as follows:-

- Product_Name
- Coles_Price
- Woolworths_Price

Visualisation Process:-

- Firstly, while collecting the data from online we have compared the same product Prices on both the Coles and Woolworths website.
- Then, we have summarised the data to check the difference of Prices based on mean,median,mode,1st Quartile,2nd Quartile,3rd Quartile.For mean of both Coles and Woolworthsthere is no significant difference between the prices of both the products.
- Boxplot identifies the possible outliers from both the Coles and Woolworths.
- After removing the outliers, further investigation is carried upon through

1. Histogram - for price differences,
2. Density plot - to verify normal assumption,
3. Line plots - nad ggplot to further visualise the data.

- Finally, we have performed paired t-test as these are dependent samples under the hypothesis testing.

Observations:- * Mean difference of Coles and Woolworths depicts that there is no significant difference between the prices of both the stores whereas Woolworths has mean prices higher than the Coles.

- P-value is higher than the critical value 0.05 which means that Woolworths have prices higher than the Coles.
- The investigation draws the facts that there is no significance difference whereas Woolworths is a bit expensive than Coles.

# Load Packages and Data

- Here,neccessary packages are loaded such as:- 1.dplyr - transform and summarise the data. 2.ggplot2 - visualize the data. 3.readr - import the data. 4.knitr - preview the rmd document in html format and convert it into pdf or other formats. 5.MVN - dect the multivariate outliers and outliers to detect the outliers.

Hide

```
library(knitr)
library(MVN)
library(ggplot2)
library(magrittr)
library(dplyr)
library(readr)
library(outliers)
library(tidyr)
library(car)# package to call the qqplot function which is used to draw qq plots
library(lattice)
```

- This chunk imports the data using readr library and read_csv() function.

Hide

```
Price_wars <- read_csv("Stats.csv")
```

```
Parsed with column specification:
cols(
  Product_Name = □[31mcol_character()□[39m,
  Woolworths_Price = □[32mcol_double()□[39m,
  Coles_Price = □[32mcol_double()□[39m
)
```

Hide

```
View(Price_wars)
```

# Summary Statistics

- Summary statistics of Price_wars is calculated through summarise() providing mean,minimum,maximum,median,mode,1st Qaurtile,2nd Quartile,3rd Quartile.
- Boxplots of all the categories as well as complete dataset show that there are couple of outliers which may be due to considering random samples of the data.
- Also, mean of coles data seems to be comparitively on a lower side compared to woolworths which shows Woolworths have cheaper prices than Coles.
- After that, outliers are removed through z.scores() method.

Hide

```
Price_wars_Coles<-Price_wars$Coles_Price
Price_wars_product_name <- Price_wars$Product_Name
columns_name <- c("Store","Products_Name","Prices")
Coles_Supermarket <- data.frame("Coles_Store",Price_wars_product_name,Price_wars_Coles)
names(Coles_Supermarket) <-columns_name
Price_wars_Woolworths <-Price_wars$Woolworths_Price
Woolworths_Supermarket <- data.frame("Woolworths",Price_wars_product_name,Price_wars_Woolwort
hs)
names(Woolworths_Supermarket) <-columns_name
Giant_Stores <- rbind(Coles_Supermarket,Woolworths_Supermarket)
Giant_Stores
```

| Store<br><fctr> | Products_Name<br><fctr> | Prices<br><dbl> |
|---|---|---|
| Coles_Store | Soy Milk | 3.85 |
| Coles_Store | Flavoured milk | 2.65 |
| Coles_Store | Cage Eggs | 3.35 |
| Coles_Store | Free Range Eggs | 4.20 |
| Coles_Store | Yogurt | 3.50 |
| Coles_Store | Breakfast Cereals | 3.80 |
| Coles_Store | Oats | 5.00 |
| Coles_Store | Museli | 3.50 |
| Coles_Store | Up and GO | 14.40 |
| Coles_Store | Jam | 2.50 |

1-10 of 96 rows Previous **1** 2 3 4 5 6 … 10 Next

Hide

```
  #Coles Prices Summary
Coles_Prices_Summary<-Price_wars %>% summarise(Mean                = mean(Price_wars_Coles, n
a.rm = TRUE),
                          Median              = median(Price_wars_Coles, na.rm = TRUE),
                          Mode                = mode(Price_wars_Coles),
                         `Standard Deviation` = sd(Price_wars_Coles, na.rm = TRUE),
                         `1st Quartile`       = quantile(Price_wars_Coles, 0.25),
                         `3rd Quartile`       = quantile(Price_wars_Coles, 0.75),
                          Min                 = min(Price_wars_Coles),
                          Max                 = max(Price_wars_Coles))

kable(Coles_Prices_Summary)
```

| Mean | Median | Mode | Standard Deviation | 1st Quartile | 3rd Quartile | Min | Max |
|---|---|---|---|---|---|---|---|
| 7.216667 | 5 | numeric | 5.320228 | 3.8375 | 9.25 | 0.9 | 20 |

Hide

```
#Woolworths Prices Summary
Woolworths_Prices_Summary<-Price_wars %>% summarise(Mean                = mean(Price_wars_Woo
lworths, na.rm = TRUE),
                           Median              = median(Price_wars_Woolworths, na.rm = TRUE),
                           Mode                = mode(Price_wars_Woolworths),
                          `Standard Deviation` = sd(Price_wars_Woolworths, na.rm = TRUE),
                          `1st Quartile`       = quantile(Price_wars_Woolworths, 0.25),
                          `3rd Quartile`       = quantile(Price_wars_Woolworths, 0.75),
                           Min                 = min(Price_wars_Woolworths),
                           Max                 = max(Price_wars_Woolworths))
kable(Woolworths_Prices_Summary)
```

| Mean | Median | Mode | Standard Deviation | 1st Quartile | 3rd Quartile | Min | Max |
|---|---|---|---|---|---|---|---|
| 7.365833 | 6.15 | numeric | 5.229657 | 3.925 | 9 | 1.1 | 21 |

Hide

```
#Both Stores Summary based on grouping of stores.
Combined_Stores_Summary <- Giant_Stores %>%  group_by(Store) %>% summarise(Min = min(Prices,n
a.rm = TRUE),
                                          Q1 = quantile(Prices,probs = .25,na.rm = TRUE),
                                          Median = median(Prices, na.rm = TRUE),
                                          Q3 = quantile(Prices,probs = .75,na.rm = TRUE),
                                          Max = max(Prices,na.rm = TRUE),
                                          Mean = mean(Prices, na.rm = TRUE),
                                          SD = sd(Prices, na.rm = TRUE),
                                          n = n(),
                                          tcrit = round(qt(p = 0.975, df = n - 1),3),
                                          SE = round(SD/sqrt(n),3),
                                          '95% CI Lower Bound' = round(Mean - tcrit * SE,2),
                                          '95% CI Upper Bound' = round(Mean + tcrit * SE,2))

names(Combined_Stores_Summary)[9] <- "Number of Products"
names(Combined_Stores_Summary)[10] <- "Critical Values"
kable(Combined_Stores_Summary)
```

| Store | Min | Q1 | Median | Q3 | Max | Mean | SD | Number of Products | Critical Values | SE | 95% CI Lower Bound | 95% CI Upper Bound |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Coles_Store | 0.9 | 3.8375 | 5.00 | 9.25 | 20 | 7.216667 | 5.320228 | 48 | 2.012 | 0.768 | 5.67 | 8.76 |
| Woolworths | 1.1 | 3.9250 | 6.15 | 9.00 | 21 | 7.365833 | 5.229657 | 48 | 2.012 | 0.755 | 5.85 | 8.88 |

- From the further visualisation through boxplot method, we have found that there are few outliers at both the stores which may be due to random sampling.

Hide

```
boxplot(Price_wars$Woolworths_Price,Price_wars$Coles_Price,main="Box plot showing comparison
 of Prices of Coles and Woolworths",ylab="Price Range",col="green")
axis(1, at = 1:2, labels = c("Woolworths_Price", "Coles_Price"))
```


- Removing the Outliers through z.scores method

Hide

```
z.scores<-Price_wars$Woolworths_Price%>%scores(type="z")
z.scores%>%summary()
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-1.1981 -0.6579 -0.2325  0.0000  0.3125  2.6071 
```

Hide

```
z.scores<-Price_wars$Coles_Price%>%scores(type="z")
z.scores%>%summary()
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-1.1873 -0.6352 -0.4166  0.0000  0.3822  2.4028 
```

Hide

```
Price_Wars_clean_Wooli<- Price_wars$Woolworths_Price[ - which( abs(z.scores) >=3 )]
Price_Wars_clean_Coles<- Price_wars$Coles_Price[ - which( abs(z.scores) >=3 )]
```

- Showing the distribution of Price Difference

Hide

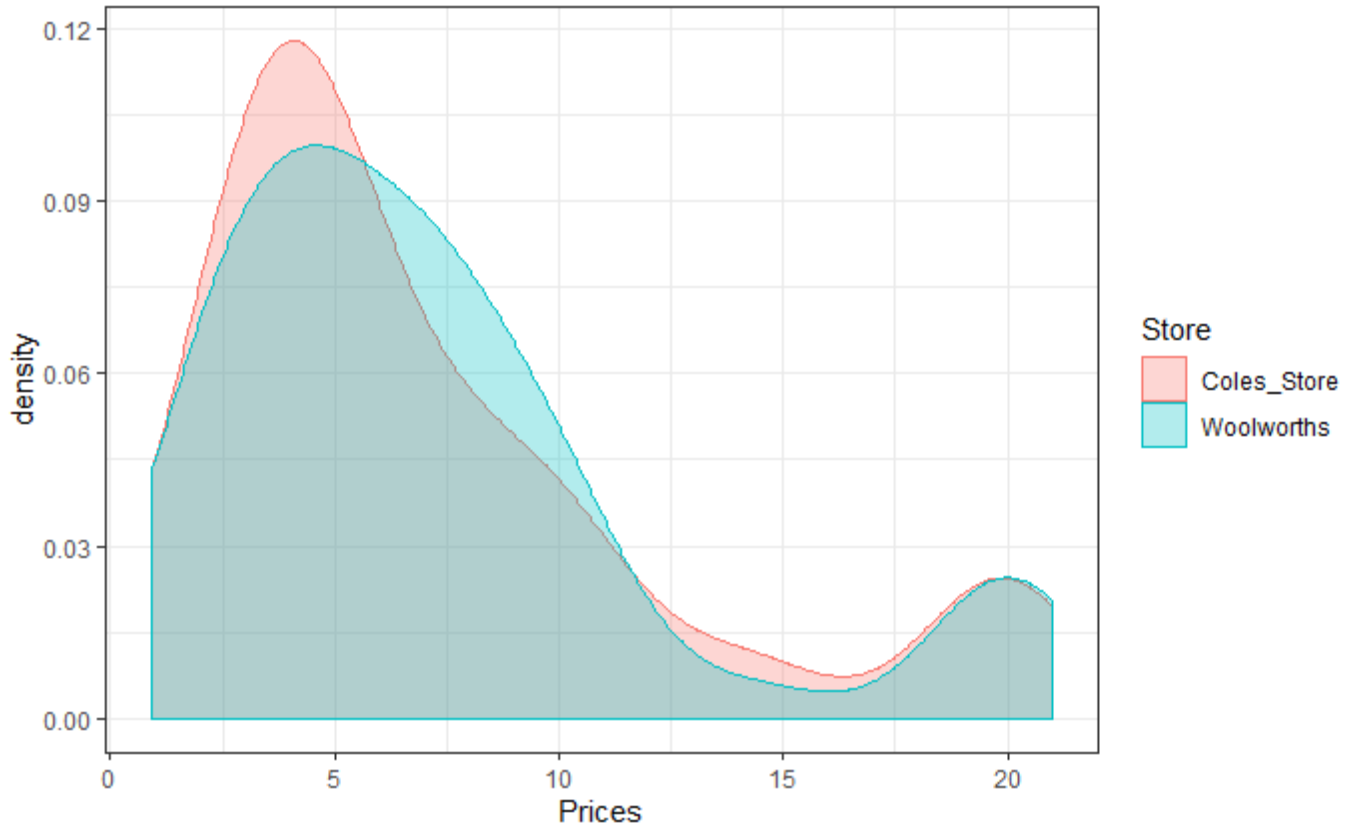
```
Price_Wars_Difference <- Price_wars %>% mutate(difference = Price_wars_Woolworths - Price_war
s_Coles )

ggplot(data=Price_Wars_Difference, aes(difference)) + geom_histogram(color = "blue", bins=18)
+
  scale_x_continuous(breaks = seq(-8,8,1)) +  xlab('Price difference') + ylab("Count_Of_Produ
cts") +
   ggtitle("Distribution of Price Difference Between Coles and Woolworths")
```


- Density Plot is plotted to depict that there is no significant visual difference observed between Coles and Woolworths prices as we can observe that Coles_Store has slightly steeper curve than Woolworths . But still, further investigation is yet to be performed.

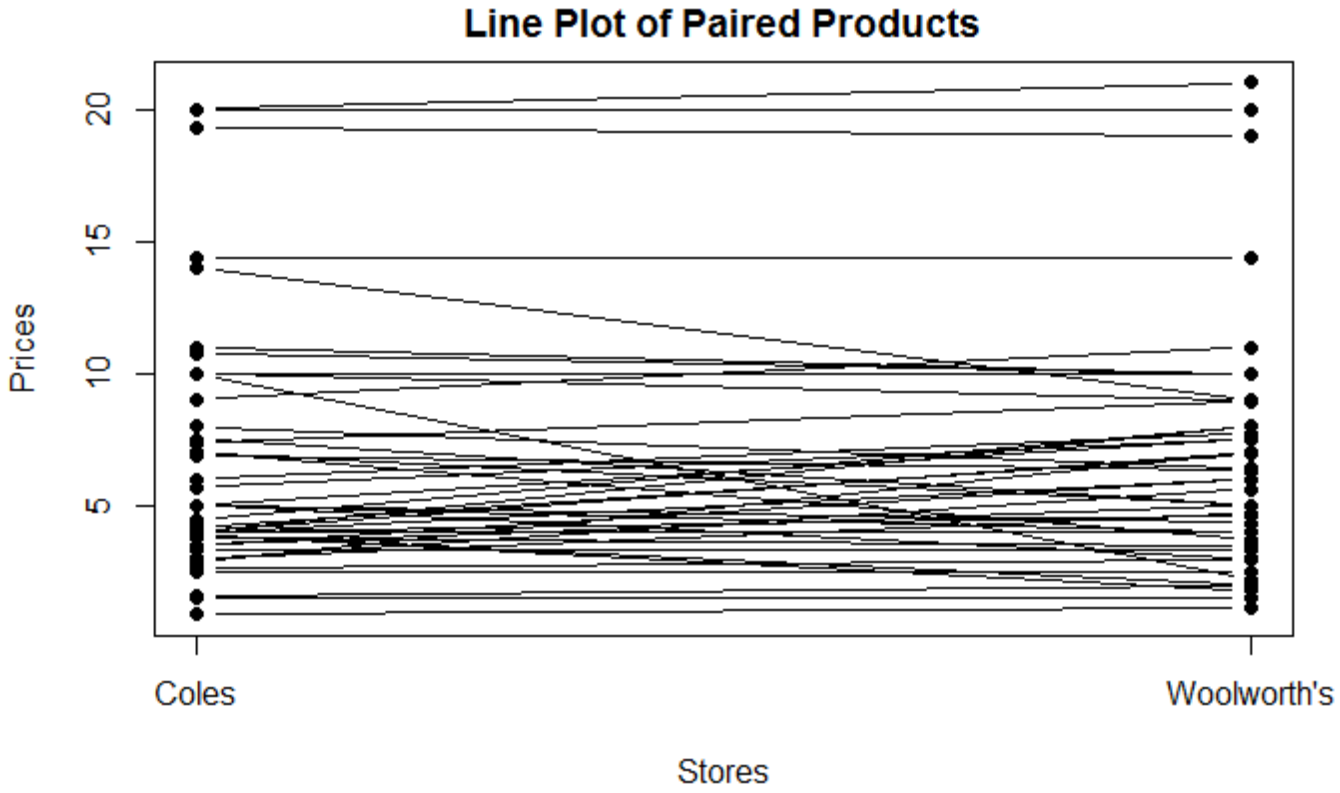
Hide

```
ggplot(Giant_Stores,aes(x=Prices)) +
  geom_density(aes(group=Store, colour=Store,fill=Store),alpha=0.3) +
  theme_bw()
```




Hide

```
matplot(t(data.frame(Price_wars$Coles_Price, Price_wars$Woolworths_Price)),
  type = "b",
  pch = 19,
  col = 1,
  lty = 1,
  xlab = "Stores",
  ylab = "Prices",
  xaxt = "n",
  main = "Line Plot of Paired Products"
  )
axis(1, at = 1:2, labels = c("Coles", "Woolworth's"))
```




Hide

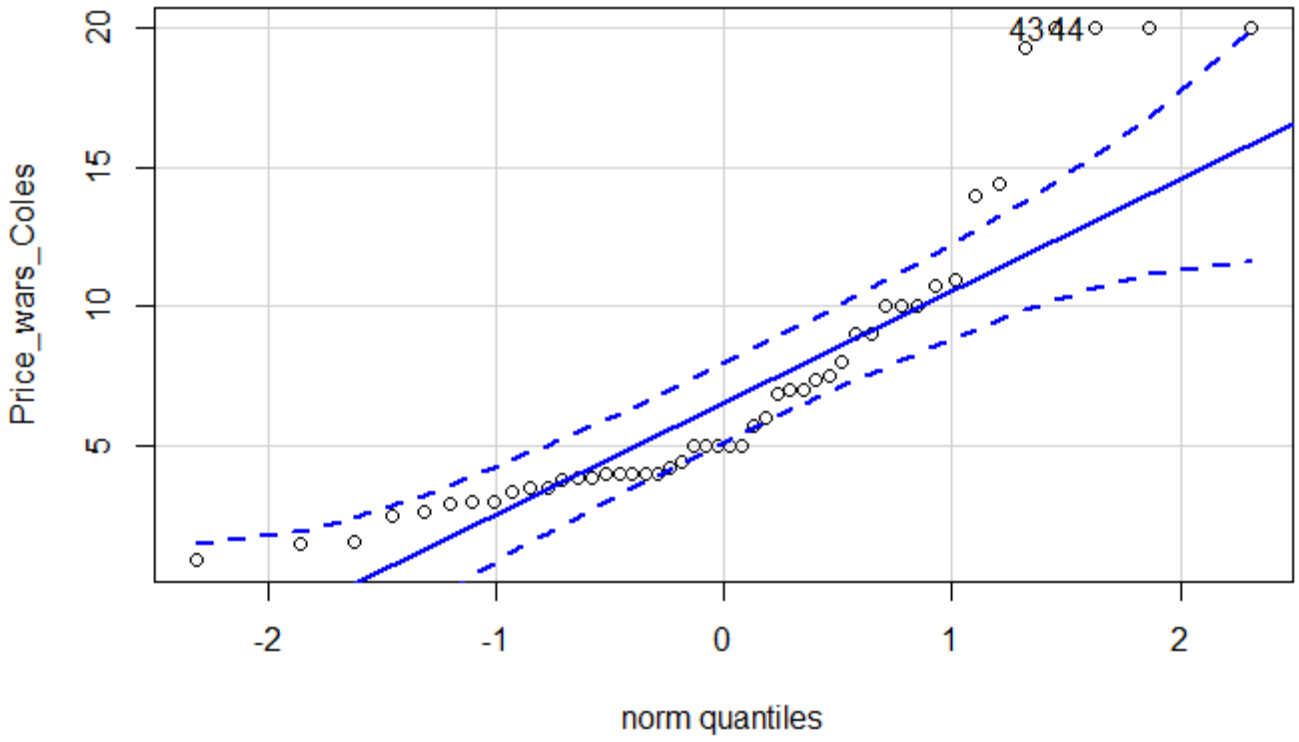
```
Giant_Stores %>% histogram(~Prices | Store, data = .,layout=c(1,2), main= 'Comparison of Pric
e Distribution')
```

- Q-Q Plot to check for the normality distribution of Prices of Coles Store

Hide

```
qqPlot(Price_wars_Coles,dist="norm")
```

```
[1] 43 44
```

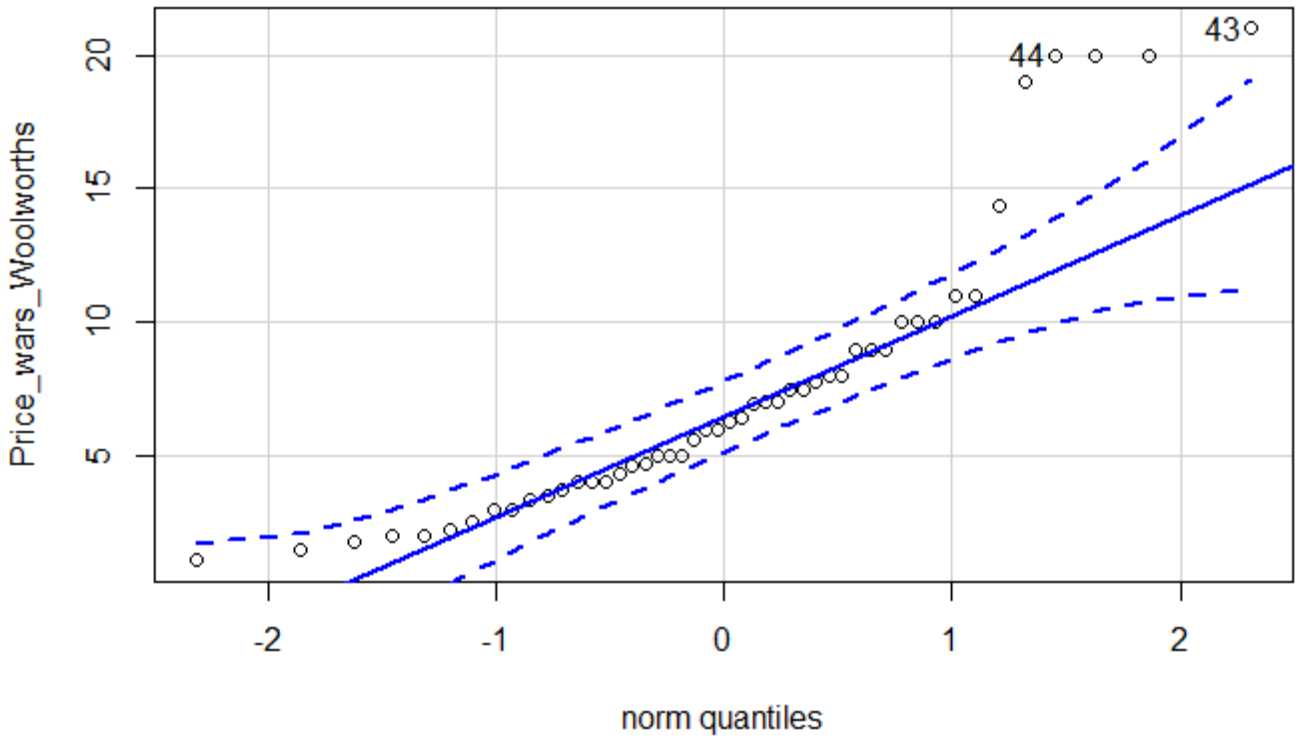



- Q-Q Plot to check for the normality distribution of Prices of Woolworths Store

Hide

```
qqPlot(Price_wars_Woolworths,dist="norm")
```

```
[1] 43 44
```



# Hypothesis Test

- To perform the Hypothesis test, we have used Welch Two-sample t-test for comparing the difference between price rate of the products from Coles and Woolworths for this report.
- The two sample t-test assumes the price samples collected from Coles and Woolworths which are independent of each other.
- The default confidence level is considered for the test is 95%. The significance level thus becomes 0.05 which is accepted as a standard error in order to prove our Null hypothesis and we have considered mu=0 in each case.
- Thus, for Null Hypothesis, we assume both the Coles and Woolworths supermarkets have same prices.
- Then, we have performed paired samples test to spot out the differences between costs of the products from the supermarkets.
- A paired samples test was chosen to analyze this as the products in each group.
- The 0.05 level of significance was used which is taken as critical value.
- As our sample size was 49 we proceeded with the t-test, assuming normality of variances due to the Central Limit Theorem.

Hide

```
t.test(Price_wars$Woolworths_Price,Price_wars$Coles_Price,data=Price_wars,mu=0,var.equal = TR
UE,paired=TRUE,
  alternative = "two.sided")
```

```
    Paired t-test

data:  Price_wars$Woolworths_Price and Price_wars$Coles_Price
t = 0.46266, df = 47, p-value = 0.6457
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4994408  0.7977742
sample estimates:
mean of the differences
              0.1491667
```

Hide

```
t.test(Price_Wars_Difference$difference,
       mu = 0,
       conf.level = 0.95,
       alternative = "two.sided")
```

```
    One Sample t-test

data:  Price_Wars_Difference$difference
t = 0.46266, df = 47, p-value = 0.6457
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.4994408  0.7977742
sample estimates:
mean of x
0.1491667
```

# Interpretation

- By doing analysis on the whole dataset containing different categories and products that we collected, there is not enough evidence to prove that either one of the supermarket is costly or cheaper.
- Considering the dataset that we have collected through boxplot we can determine Woolworths is slighty expensive than Coles.
- When t-test is performed on the complete dataset without any classifications, it is found that Woolworths is slightly expensive than Coles.
- These results are based on the dataset that we collected for the investigation and it may vary if the categories and products are increased for the report.
- Also the differences in the price of these two supermarkets is not huge.
- Although, mean difference was found to be 0.1491667 which lies between Confidence Interval(-0.378 to 0.664).

Deriving the fact from above conclusions, it can be said that our test result was not statistically significant, or it failed to reject the null hypothises(H0), implying the price differences between the products of Coles and Woolsworths is equal to 0, or negligibly different.

# Discussion

- As per our analysis, we found out that Coles is cheaper than Woolworths.
- However, our analysis was limited to the random samples we collected.
- The buyer should compare the products categorically as there might be some categories where Woolworths is cheaper than Coles and vice versa, to get the cheapest product in the market.
- The strength of our analysis stands for these common purchase categories that Coles shall be preferred over Woolworths for cheaper rates and the limitation is the size of our dataset and limited number of categories in the data due to which we cannot compare the two supermarkets in every category.
- In future, we would like to increase the amount of data collected to carry out this investigation and find out if there are any changes in the conclusions of the research.
- We may change the way we collected the data as we focused on random sampling rather than on stratified sampling.
- 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 amount of data, the analysis is going to be more significant.
- Limitations of this investigation was the use of the online store in selecting products rather than buying them from the store, where there may be discrepancies between prices between the two.