Code ▾

MATH1324 Assignment 2

Supermarket Price Wars

Group/Individual Details

- Vikas Virani (s3715555)
- Vishwa Gandhi (s3714805)
- Jigar Mangukiya (s3715807)

Executive Statement

```
  Aim of this study is to confirm if there is significant statistical evidence that one of th
e two supermarkets operating in Australia, Coles & Woolworths, is cheaper than the other. The
sample to conduct the investigation consist of 48 products sold by both supermarkets, with 4
variables, namely Product, Size, Woolworths Price & Coles Price.

  (The data is acquired from - https://www.finder.com.au/grocery-price-comparison.)

  We have provided major summary statistic with a grouped line chart to provide initial visua
lization of data set. We have also plotted a qqplot to confirm if the data is normally distri
buted or not. However, the sample size is large (n = 48), so we can skip the qqplot, as CLT w
ill imply that sampling distribution will be normal for large number of samples.

  To confirm the statistical evidence that shows either of the supermarket is cheaper than th
e other, we have used "paired sample t-test".

  Analyzing the results from t-test, they were found to be not statistically significant to r
eject our original hypothesis. That led us to conclude the study in favor of the statement,
"There is no significant price difference in the prices of two supermarkets."
```

Load Packages and Data

Hide

```
# This is a chunk where you can load the necessary data and packages required to reproduce th
e report
# You should also include your code required to prepare your data for analysis.
library(readr)
library(magrittr)
library(dplyr)
library(stats)
library(car)
library(ggplot2)
library(tidyr)
Price_Wars_Products <- read_csv("Product_Prices_Coles_WoolWorths.csv")
```

Hide

```
head(Price_Wars_Products)
```

Product <chr>	Size <chr>	Woolworths Price <dbl>	Coles Price <dbl>
Granny Smith Apples	1kg	2.0	4.0
Fresh tomatoes	500g	7.9	7.5
Watermelon	Whole	8.4	11.2
Cucumber	1 whole	2.0	2.0
Red potato washed	1kg	4.0	4.0
Red-tipped bananas	1kg	5.0	5.0

6 rows

Summary Statistics

As “Coles Price” variable is character, we first convert it into numeric to summarize the variable.

As from the QQ Plot, The differences appear to have some values outside of 95% CI for normal quantiles, but as sample size for each group is greater than 30(i.e.48 products in each) we can proceed even if the normality assumption is violated.

Hide

```
# This is a chunk for your summary statistics and visualization  code
Price_Wars_Products$`Coles Price` <- Price_Wars_Products$`Coles Price` %>% as.numeric()
Price_Wars_Products %>% summarise(Min = min(`Coles Price`,na.rm = TRUE),
                                             Q1 = quantile(`Coles Price`,probs = .25,na.rm = 
TRUE),
                                             Median = median(`Coles Price`, na.rm = TRUE),
                                             Q3 = quantile(`Coles Price`,probs = .75,na.rm = 
TRUE),
                                             Max = max(`Coles Price`,na.rm = TRUE),
                                             Mean = mean(`Coles Price`, na.rm = TRUE),
                                             SD = sd(`Coles Price`, na.rm = TRUE),
                                             n = n(),
                                             Missing = sum(is.na(`Coles Price`)))
```

Min <dbl>	Q1 <dbl>	Median <dbl>	Q3 <dbl>	Max <dbl>	Mean <dbl>	SD <dbl>	n <int>	Missing <int>
1.1	3.41	4.21	5.5975	11.2	4.595417	2.030607	48	0

1 row

Hide

```
Price_Wars_Products %>% summarise(Min = min(`Woolworths Price`,na.rm = TRUE),
                                             Q1 = quantile(`Woolworths Price`,probs = .25,na.
rm = TRUE),
                                             Median = median(`Woolworths Price`, na.rm = TRUE
),
                                             Q3 = quantile(`Woolworths Price`,probs = .75,na.
rm = TRUE),
                                             Max = max(`Woolworths Price`,na.rm = TRUE),
                                             Mean = mean(`Woolworths Price`, na.rm = TRUE),
                                             SD = sd(`Woolworths Price`, na.rm = TRUE),
                                             n = n(),
                                             Missing = sum(is.na(`Woolworths Price`)))
```

Min <dbl>	Q1 <dbl>	Median <dbl>	Q3 <dbl>	Max <dbl>	Mean <dbl>	SD <dbl>	n <int>	Missing <int>
1	2.9625	4	5.075	10	4.335208	2.067954	48	0

1 row

Hide

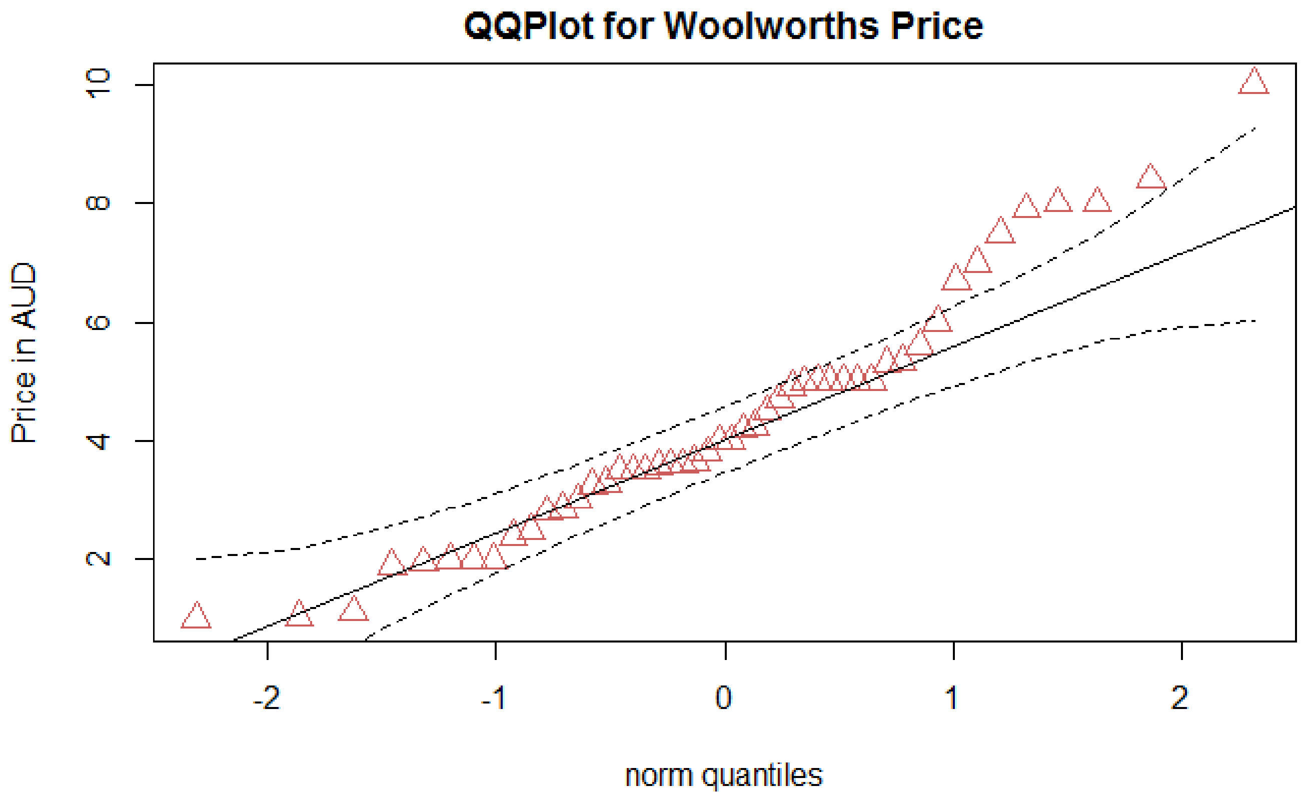
```
Price_Wars_Products <- Price_Wars_Products %>% mutate(Difference=`Woolworths Price` - `Coles
Price`)
Price_Wars_Products %>% summarise(Min = min(Difference,na.rm = TRUE),
                                             Q1 = quantile(Difference,probs = .25,na.rm = TRU
E),
                                             Median = median(Difference, na.rm = TRUE),
                                             Q3 = quantile(Difference,probs = .75,na.rm = TRU
E),
                                             Max = max(Difference,na.rm = TRUE),
                                             Mean = mean(Difference, na.rm = TRUE),
                                             SD = sd(Difference, na.rm = TRUE),
                                             n = n(),
                                             Missing = sum(is.na(Difference)))
```

Min <dbl>	Q1 <dbl>	Median <dbl>	Q3 <dbl>	M... <dbl>	Mean <dbl>	SD <dbl>	n <int>	Missing <int>
-2.8	-0.6075	-0.15	0.0025	2.5	-0.2602083	1.074209	48	0

1 row

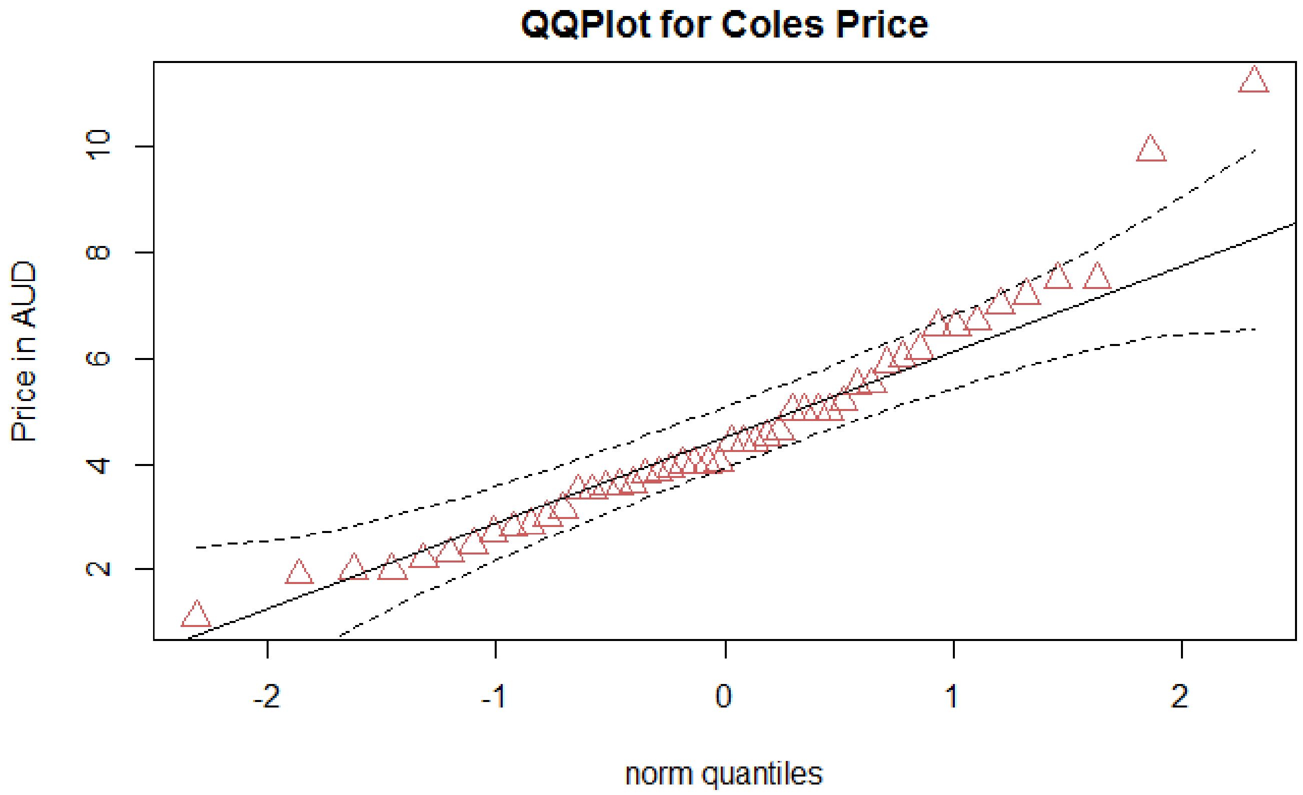
Hide

```
Price_Wars_Products$`Woolworths Price` %>% qqPlot(dist="norm",main = "QQPlot for Woolworths P
rice",col = "indianred",lwd = 1,col.lines = "black",pch = 2,ylab="Price in AUD",grid = FALSE,
cex = 1.5)
```



Hide

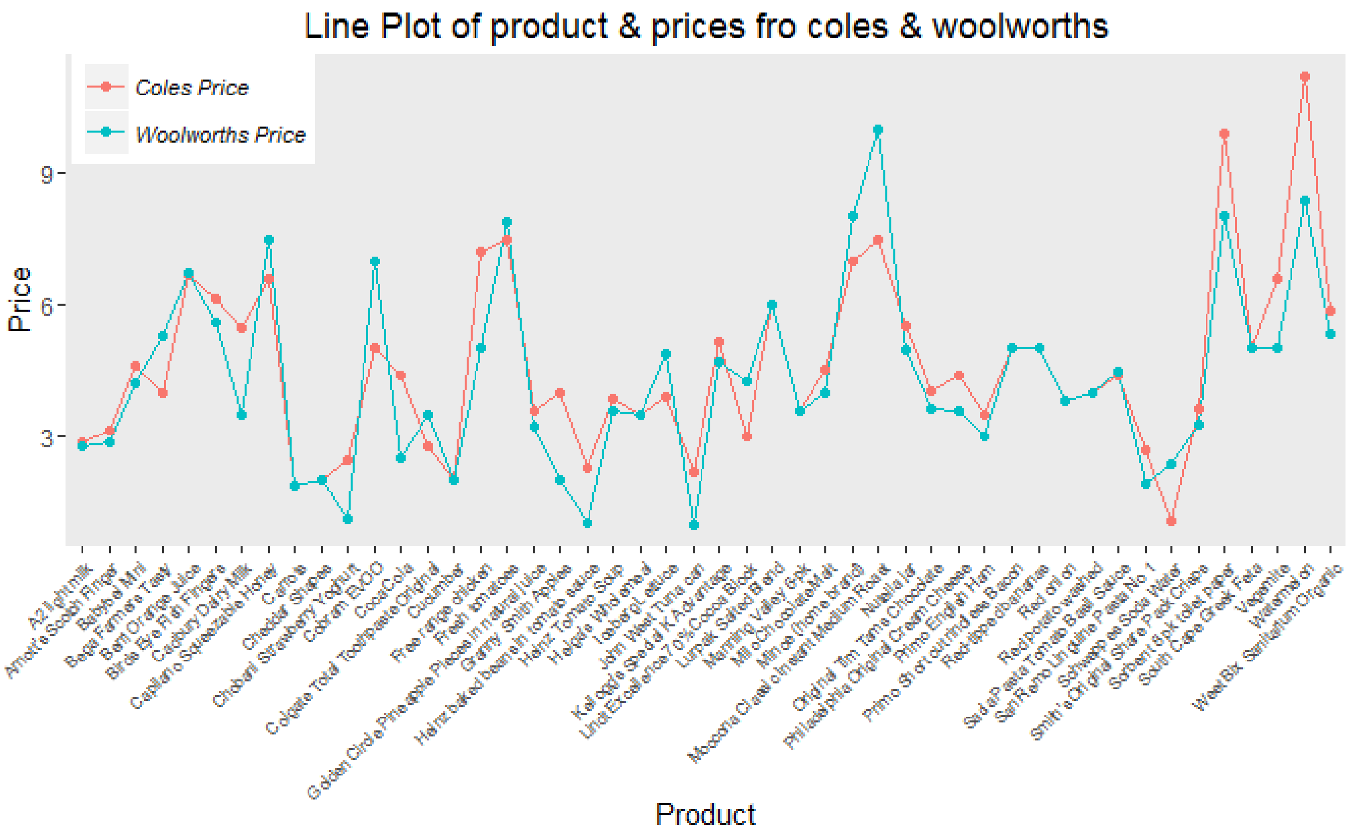
```
Price_Wars_Products$`Coles Price` %>% qqPlot(dist="norm",main = "QQPlot for Coles Price",col
 = "indianred",lwd = 1,col.lines = "black",pch = 2,ylab="Price in AUD",grid = FALSE,cex = 1.5
)
```



Hide

```
price_gathred <- Price_Wars_Products %>% gather(`Coles Price`,`Woolworths Price`,key = "Stor
e",value = "Price",convert = TRUE,na.rm = TRUE)
ggplot(data=price_gathred,
       aes(x=Product, y=Price,group=Store,colour=Store)) +
       geom_line() +geom_point() + ggtitle("Line Plot of product & prices fro coles & woolwor
ths")+

  theme(axis.text.x=element_text(size=6, angle=45, vjust=1, hjust=1),
        panel.grid.major.x=element_blank(),
          panel.grid.minor.x=element_blank(),
          panel.grid.minor.y=element_blank(),
          panel.grid.major.y=element_blank()) +
    theme(legend.text = element_text(size=8, face="italic"),
          legend.title = element_blank(),
          legend.position=c(0.1, 0.9))
```



Hide

```
NA
```

Hypothesis Test

As the sample is taken from different populations with Matched Members(i.e Matched Products), it is a dependent sample and hence we will use Paired Sample t-test to check statistically significant mean change in products' price, assuming a two-tailed test with Significance level "ALPHA"=0.05.

Hide

```
# This is a chunk for your hypothesis testing code.
t_test_result <- t.test(Price_Wars_Products$`Woolworths Price`, Price_Wars_Products$`Coles Pr
ice`,
       paired = TRUE,
       alternative = "two.sided")
t_test_result
```

```


    Paired t-test

data:  Price_Wars_Products$`Woolworths Price` and Price_Wars_Products$`Coles Price`
t = -1.6782, df = 47, p-value = 0.09994
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.57212618  0.05170952
sample estimates:
mean of the differences 
             -0.2602083 
```

Hide

```
# two-tailed p-value for the t-value of sample
p_val <- 2*pt(q = -1.6782, df = 47)
cat("Confirming significance with p-value :","\n","P Value for two sided t test: ",p_val, ",
 Which is greater than alpha (0.05).\n > Hence we fail to reject the H0 Hypothesis. \n")
```

```
Confirming significance with p-value : 
 P Value for two sided t test:  0.09994633 , Which is greater than alpha (0.05).
 > Hence we fail to reject the H0 Hypothesis. 
```

Hide

```
cat("\n\n Confirming significance with CI :","\n","95% CI region for two sided t test: [ ",t_
test_result$conf.int[[1]]," - ",t_test_result$conf.int[[2]], "], Which is includes the differ
ence of means (0).\n > Hence we fail to reject the H0 Hypothesis.")
```

```



 Confirming significance with CI : 
 95% CI region for two sided t test: [  -0.5721262  -  0.05170952 ], Which is includes the di
fference of means (0).
 > Hence we fail to reject the H0 Hypothesis.
```

Interpretation

```
A paired-samples t-test was used to test for a significance of difference of means of Product
s price from Woolworths and Coles. The difference of means of Products price was found to be
-0.260 (SD = 1.074). Visual inspection of the Q-Q plot of the difference scores suggested tha
t the data weren't approximately normally distributed, but as sample size is greater than 30,
we can proceed without being concerned with normality assumption.

The paired-samples t-test does not found a statistically significant mean difference between
Products price from Woolworths and Coles, t(df=47) = -1.678(t* = 2.012), p > .05 (i.e. 0.099
9), 95% CI [-0.572, 0.0517].

The difference of mean (0) is with in the given 95% CI and the p-value is 0.0999 which is gre
ater than alpha 0.05.
```

Discussion

```
As can be seen from Paired sample t-test results, there is no statistically significant diffe
rence between mean Products price of Woolworths and Coles and hence neither of the supermarke
t is cheap compared to another. But, the probability was found to be 0.0999 for 95% CI.

However, if the significance level was taken as 10% and we have found 90% CI than the result
could have been different as P(0.0999 < 0.1) and 90% CI would not cover mean difference of ze
ro for the products price and as a result, Woolworths Products price have been cheaper than C
oles. (This calculation is not published in this document.)

The strength of the investigation is the sample size was large(more than 30), so even if it d
oes not hold normality assumption, we can proceed with the test. Apart from that, matching pr
oducts and same quantities and brands of products are being compared which increases the accu
racy of the investigation.

The limitations of the investigation is that, the sample we used to carry out investigation d
oes not capture the gist of the huge portfolio of products offered by both supermarkets. The
sample doesn't cater to different categories of the products which can have an impact on the
investigation. Sample might not be representative sample.
The study was performed mostly on a generic set of products available at both supermarkets, H
owever, the store owned branded products (e.g. coles milk & woolworths milk etc.), which occu
py a large shelf space in both stores, can be a true point of comparison, but it cannot be in
cluded in this study due to the limited scope of the document.

Viable improvement to the study can be, use of larger sample to cover almost all possible mat
ched products to have an accurate set of results.

But with the current set of data and results, it can be inferred that there is no significant
difference in the prices of products in two supermarkets.
```