Supermarket Price Wars

Group Details

- Anooja Mathew (s3767921)
- Lipika Sharma (s3764073)
- Bhavy Shukla (s3776464)

Executive Statement

The investigation was performed to analyse the prices of Coles and Woolworths and statistically prove which one of them is more expensive. The prices of products from a stratified sample set of 90 products belonging to different categories like grocery, household, heath and beauty were collected from Coles and Woolworths. The paired t-test hypothesis was used to derive the conclusion. The mean value of the differences in prices between the stores was taken as 0 for the null hypothesis. The t-statistic was then compared to a two-tailed t-critical value $t*$ and a statistically significant difference was found between the prices. This led to rejecting the null hypothesis and accepting the alternate hypothesis.

Load Packages and Data

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## 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(qqtest)
```

```
## Warning: package 'qqtest' was built under R version 3.5.3
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.3
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

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

library(granova)

## Warning: package 'granova' was built under R version 3.5.3

library(readxl)

## Warning: package 'readxl' was built under R version 3.5.3

Assignment_Data <- read_excel("D:/Data Science/Semester
1/Statistics/Assigments/Assignment 2/Assignment_Data.xlsx")
View(Assignment_Data)
```

Summary Statistics

Summary for Coles prices

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

## # A tibble: 1 x 9
##     Min    Q1 Median    Q3   Max  Mean    SD     n Missing
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
## 1   0.6  3.52      5   7.5    25  6.29  4.49    90       0
```

Summary for Woolworths prices

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

## # A tibble: 1 x 9
##     Min    Q1 Median    Q3   Max  Mean    SD     n Missing
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
## 1  0.63   3.2   4.95  7.38  19.0  5.98  3.96    90       0
```

Summary for Difference between Coles and Woolworth prices

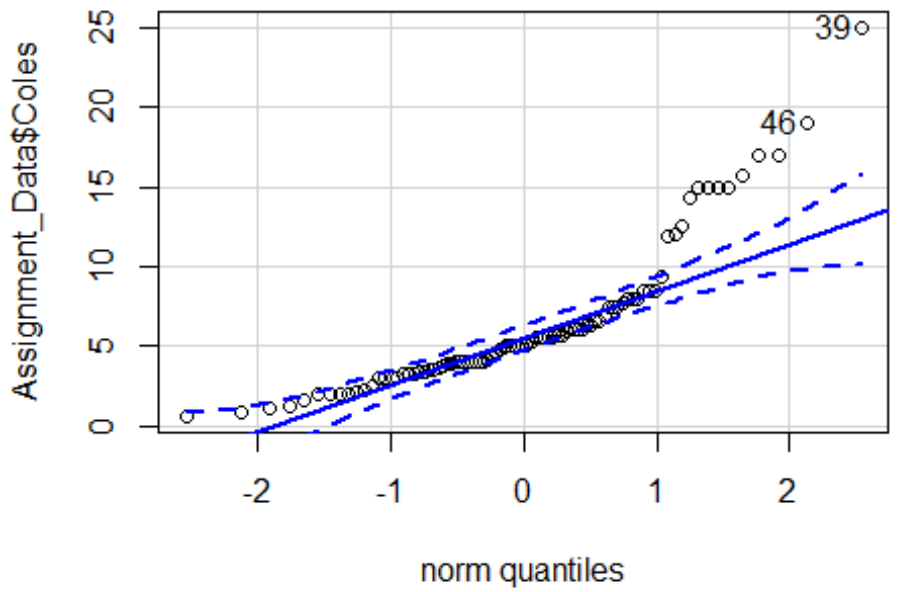
```
Assignment_Data <- Assignment_Data %>% mutate(Price_Difference = Coles -
Woolworths)

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

```
## # A tibble: 1 x 9
##     Min    Q1 Median    Q3   Max  Mean    SD     n Missing
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
## 1    -2     0      0  0.25  8.33 0.311  1.39    90       0
```

Q-Q plots for Coles

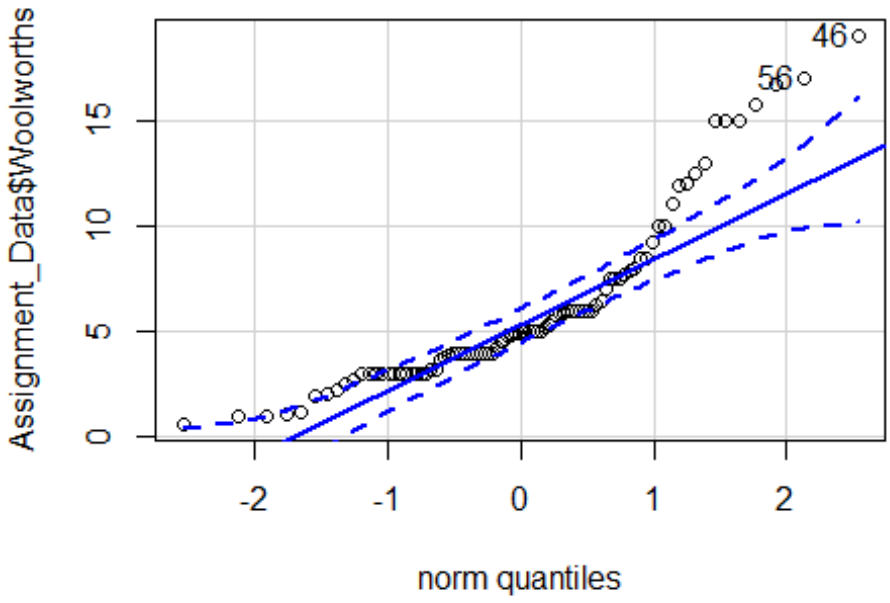
```
qqPlot(Assignment_Data$Coles,dist="norm");
```




```
## [1] 39 46
```

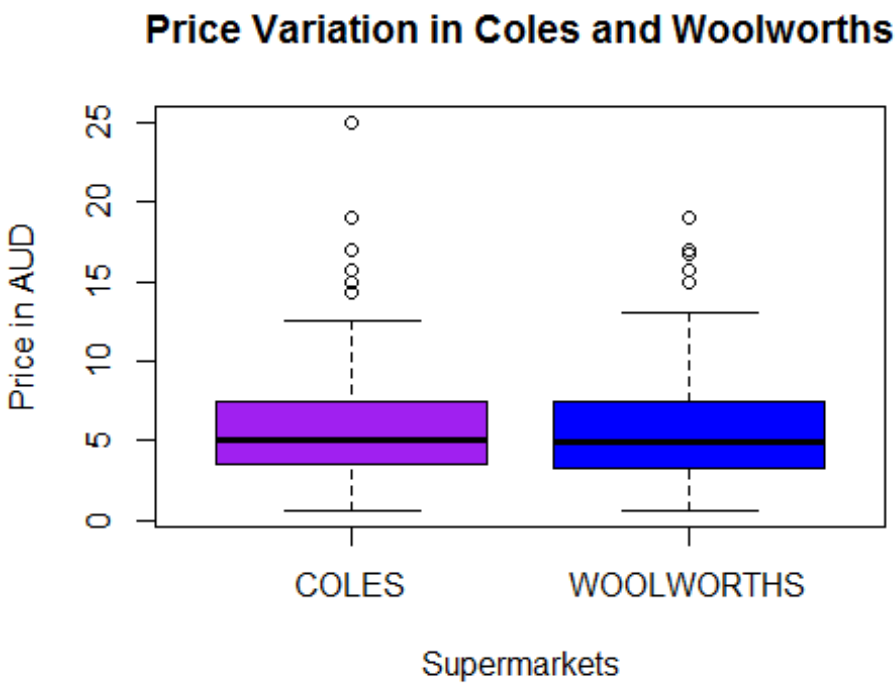
Q-Q plots for Woolworths

```r
qqPlot(Assignment_Data$Woolworths, dist="norm")
```




```
## [1] 46 56
```

```r
names=c("COLES", "WOOLWORTHS")
Assignment_Data %>% boxplot(Assignment_Data$Coles,
Assignment_Data$Woolworths, names=names, data = .,
main="Price Variation in Coles and Woolworths",
xlab="Supermarkets", ylab="Price in AUD", col=c("Purple","Blue"))
```



Hypothesis Test

The statistical hypotheses for the paired-samples t-test are as follows:

Null Hypothesis: The difference between the mean values of products from Coles and Woolworths is 0 i.e. H0:μ1−μ2=0

Alternate Hypothesis: The difference between the mean values of products from Coles and Woolworths is not 0 i.e. H0:μ1−μ2 ≠ 0

Paired t -test

```
 t.test(Assignment_Data$Coles, Assignment_Data$Woolworths,
       paired = TRUE,
       alternative = "two.sided")

##
##  Paired t-test
##
## data:  Assignment_Data$Coles and Assignment_Data$Woolworths
## t = 2.1225, df = 89, p-value = 0.03657
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01986124 0.60213876
## sample estimates:
## mean of the differences
##                   0.311
```

One sample t-test

```
t.test(Assignment_Data$Price_Difference, conf.level = 0.95, alternative =
"two.sided")
```

```
##
##  One Sample t-test
##
## data:  Assignment_Data$Price_Difference
## t = 2.1225, df = 89, p-value = 0.03657
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.01986124 0.60213876
## sample estimates:
## mean of x
##     0.311
```

Critical value t* for the paired t-test are ±1.98 as below. As t=2.12 is more extreme than 1.98, H0 should be rejected. There was a statistically significant mean difference between Coles and Woolworths.

```
qt(p = 0.025, df = 89)
```

```
## [1] -1.986979
```

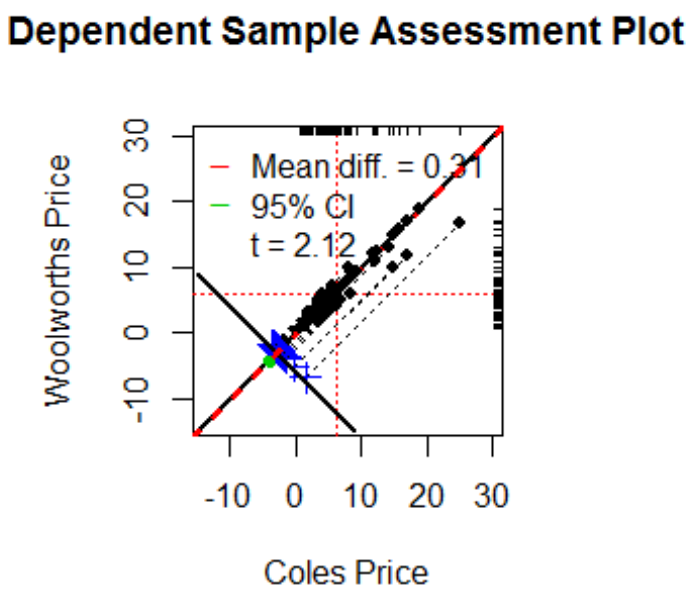
The two-tailed p-value can be calculated as below which rounds to p <.037 reported in the paired samples t-test. As p < .05, we reject H0. There was a statistically significant mean difference.:

```
2*pt(q = -2.12, df = 89)
```

```
## [1] 0.03678932
```

Paired Samples t-test Visualisation

```
granova.ds(
  data.frame(Assignment_Data$Coles, Assignment_Data$Woolworths),
  xlab = "Coles Price",
  ylab = "Woolworths Price"
  )
```




```
##              Summary Stats
## n                   90.000
## mean(x)              6.292
## mean(y)              5.981
## mean(D=x-y)          0.311
## SD(D)                1.390
## ES(D)                0.224
## r(x,y)               0.954
## r(x+y,d)             0.387
## LL 95%CI             0.020
## UL 95%CI             0.602
## t(D-bar)             2.123
## df.t                89.000
## pval.t               0.037
```

Interpretation

A paired-samples t-test was used to test for a significant mean difference between the prices for Coles and Woolworths. The mean difference for the stores was found to be 0.311 (SD = 1.39). Visual inspection of the Q-Q plots suggested that the data were not approximately normally distributed. The paired-samples t-test found a statistically significant mean difference between stores, t(df=89)=2.12, p<.03, 95% [-0.02, 0.60]. Price range for products of Woolworths were found to be significantly less as compared to Coles.

Discussion

The investigation above helps us to conclude that Woolworths is cheaper than Coles. However, there were limitations faced with respect to the sample data collected. The sample size taken is much less than the actual sample size i.e. only 90 products were considered for the analysis instead of a larger product range. Another major limitation would be the frequent change in prices of both the stores. The analysis would have been more productive if a larger sample size consisting of a wider proportion of the products were collected .

References:

https://shop.coles.com.au/a/a-national/everything/browse
https://www.woolworths.com.au/shop/discover/shopping-online