MATH1324 Assignment 2

Code

Supermarket Price Wars

Group/Individual Details

- Rafeed Sultaan (s3763175)
- Vishal Beniwal (s3759790)
- Jewel James (s3763905)

Executive Statement

The main goal of the report was to investigate to check if there was a price difference between two Australian superstores: Coles and Woolworths. The approach of the investigation included taking two sets of matching sample products from each of the superstores. The sample size of the products was 80. We used a stratified sampling approach for random sampling by dividing the products into sub-categories: Fruits, Breakfast, Frozen, Vegetable, Snacks, Dairy, and Raw. In each of the sub-categories, we used simple random sampling (SRS) approach to randomly select a small sample which was representative of each of the sub-categories.

To collect the data, both the stores were physically visited where we created a basket for each sub-category. The items inside the basket were randomly selected from the available items across a particular category with the intention of matching the selected products from both the stores. During the item selection, to make sure the availability of the item, we also considered the popularity factor of the brand for a particular item in order to increase the probability of finding that same item in both the stores.

After the data preparation process, We performed a paired sample t-test for both of the stores. The null hypothesis was that the product price was the same for both stores. After conducting the hypothesis test, we found a statistically significant difference in the product price of both stores. As we found the significant difference, we performed another hypothesis test with the intent to determine which superstore is cheaper. The result of the second hypothesis test provided sufficient evidence to support that the product price of Coles is cheaper than the Woolworths.

Load Packages and Data

Hide

```
#loaded required libraries
library(readr)
library(magrittr)
library(dplyr)
library(qqtest)
library(car)
library(ggplot2)
library(granova)
```

Summary Statistics

The mean price of Coles items was found to be 6.1, whereas the mean price was found to be 6.4 for Woolworths. The minimum and maximum item price for Coles is 0.43 & 28.55, on the other hand, 0.45 and 35 for the Woolworths

The Q-Q plot was plotted to check whether the sampled data satisfied the normal distribution of the population. In the QQ plot, if the data are normally distributed, the data points will fall close to the diagonal line. In the QQ plot, we have identified the 'S' shaped of the data points in the graphs for both of the stores, which signifies the non-normality of the sample data. Fortunately, according to Central Limit Theorem the large sample size, 80> 30, the normality assumption is justified.

By comparing the boxplots of both stores, the median product price of Coles is 2.87 which was less than the median product price of Woolworths which was 3.17. This means that the product price of 50% of the sampled products collected from Coles was cheaper than the sampled products collected from Woolworths. the interquartile range of Coles samples is 8.22 which was slightly less than the interquartile range of Woolworth's sample which is 7.69. This information indicates that the spread of product price was less for Coles compared to Woolworths. Both boxplots contained a lot of outliers because there were some products which had extreme prices compared to 50% of products collected from both stores. Finally, from the boxplots, we can see that the distribution price of both of the samples collected are right-skewed.

Hide

```
#Descriptive Statistics for Coles
Store %>%  summarise(Min = min(ColesPrice,na.rm = TRUE),
                                    Q1 = quantile(ColesPrice,probs = .25,na.rm = TR
UE),
                                    Median = median(ColesPrice, na.rm = TRUE),
                                    Q3 = quantile(ColesPrice,probs = .75,na.rm = TR
UE),
                                    IQR = IQR(ColesPrice, na.rm = TRUE),
                                    Max = max(ColesPrice,na.rm = TRUE),
                                    range = Max - Min,
                                    Mean = round(mean(ColesPrice, na.rm = TRUE),1),
                                    SD = round(sd(ColesPrice, na.rm = TRUE),2)
                                    )
```

Min	Q1	Median	Q3	IQR	Max	range	Mean	SD
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0.43	1.135	2.865	9.35	8.215	28.55	28.12	6.1	6.82

1 row

Hide

```
#Descriptive Statistics for Woolworths
Store %>%  summarise(Min = min(WoolworthsPrice,na.rm = TRUE),
                                    Q1 = quantile(WoolworthsPrice,probs = .25,na.rm
= TRUE),
                                    Median = median(WoolworthsPrice, na.rm = TRUE),
                                    Q3 = quantile(WoolworthsPrice,probs = .75,na.rm
= TRUE),
                                    IQR = IQR(WoolworthsPrice, na.rm = TRUE),
                                    Max = max(WoolworthsPrice,na.rm = TRUE),
                                    range = Max - Min,
                                    Mean = round(mean(WoolworthsPrice, na.rm = TRUE
),1),
                                    SD = round(sd(WoolworthsPrice, na.rm = TRUE),2
))
```

Min	Q1	Median	Q3	IQR	Max	range	Mean	SD
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0.45	1.3575	3.165	9.0425	7.685	35	34.55	6.4	7.49

1 row

Hide

```
#Descriptive Statistics for the difference in the price of both the superstores
Store %>% summarise(Min = min(Difference,na.rm = TRUE),
                                    Q1 = quantile(Difference,probs = .25,na.rm = TR
UE),
                                    Median = median(Difference, na.rm = TRUE),
                                    Q3 = quantile(Difference,probs = .75,na.rm = TR
UE),
                                    IQR = IQR(Difference, na.rm = TRUE),
                                    Max = max(Difference,na.rm = TRUE),
                                    range = Max - Min,
                                    Mean = round(mean(Difference, na.rm = TRUE),1),
                                    SD = round(sd(Difference, na.rm = TRUE),2))
```

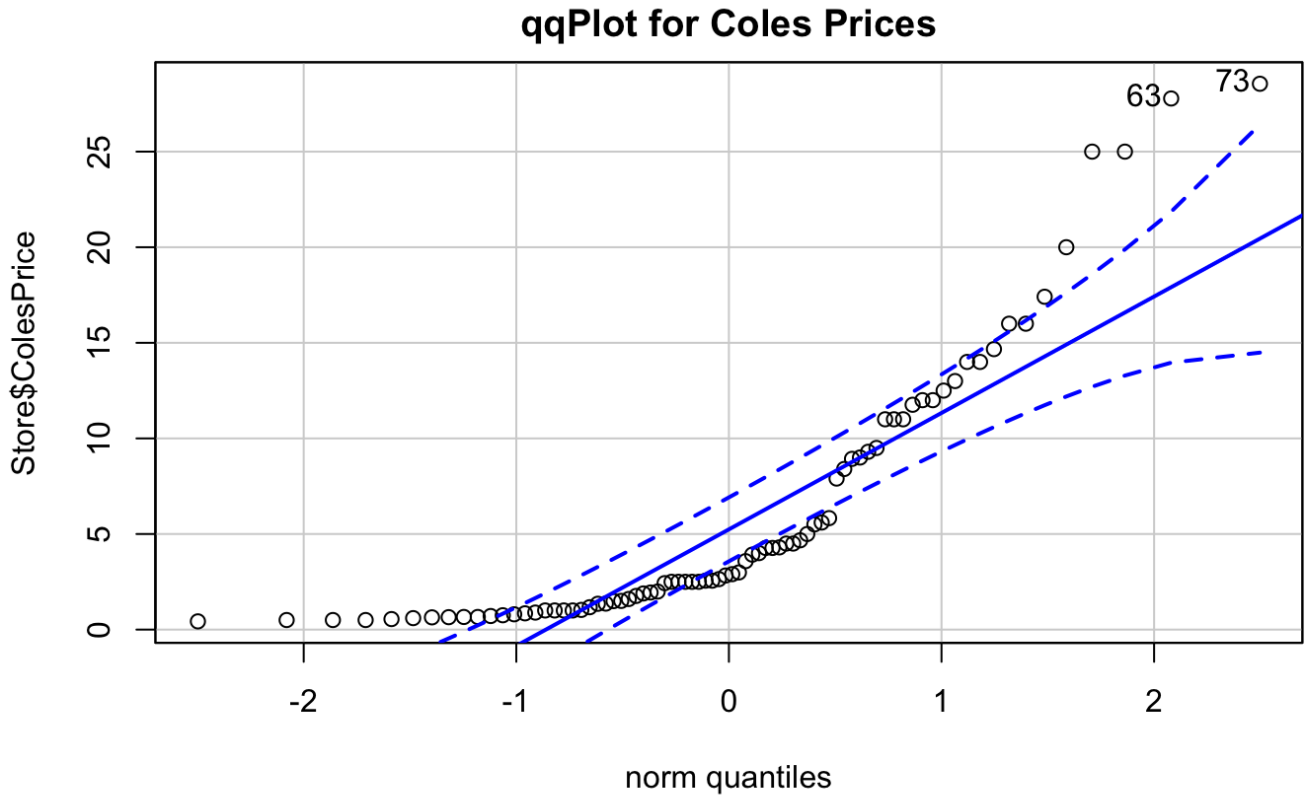
Min	Q1	Median	Q3	IQR	Max	range	Mean	SD
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
-7.22	-0.3475	0	0.0025	0.35	3	10.22	-0.3	1.5

1 row

Hide

```
#Checking the normality for sampled data points present for the coles
qqPlot(Store$ColesPrice, dist="norm", main="qqPlot for Coles Prices")
```

```
[1] 73 63
```



Hide

```
#Checking the normality for sampled data points present for the Woolworths
qqPlot(Store$WoolworthsPrice, dist="norm", main="qqPlot for Woolworths Prices")
```

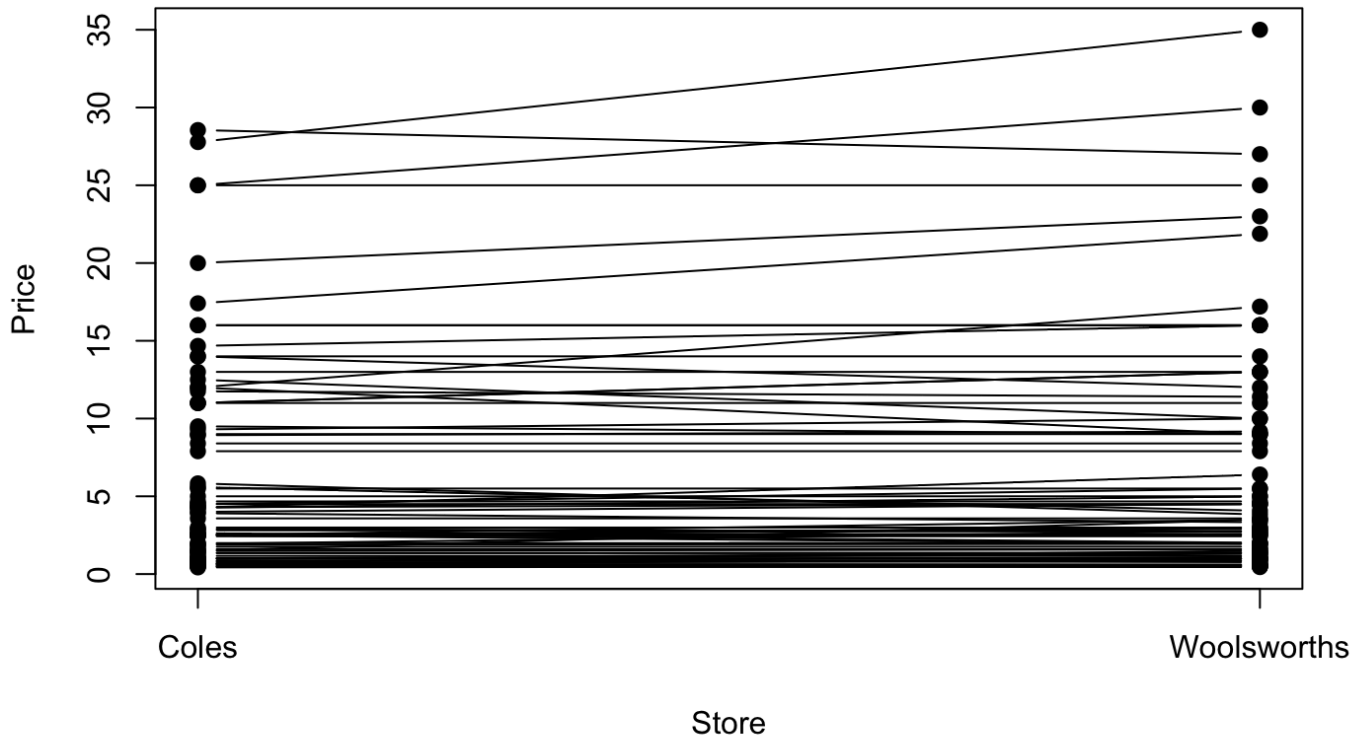
```
[1] 63  7
```

Hide

```
#Price Difference between two stores
Store %>% boxplot(Store$ColesPrice, Store$WoolworthsPrice, names=c("Coles", "Woolwort
hs"), data = .,
                  main="Boxplot : Price War Between Coles & Woolworths",
                  xlab="Store", ylab="Price", col=c("lightblue", "gold"))
```

Hide

```
matplot(t(data.frame(Store$ColesPrice, Store$WoolworthsPrice)),
  type = "b",
  pch = 19,
  col = 1,
  lty = 1,
  xlab = "Store",
  ylab = "Price",
  xaxt = "n"
  )
axis(1, at = 1:2, labels = c("Coles", "Woolsworths"))
```



Hypothesis Test

As the store items were matched, paired-samples t-test was conducted to determine whether there was a significant difference between the mean price of Coles and Woolworth's items. The 0.05 level of significance was used.

The null hypothesis (H0) of the first hypothesis test is the assumption that the product prices of both stores are equal. The alternative hypothesis (HA) of the second hypothesis test is the assumption product price of both stores are not equal. Based on the results of the first hypothesis test, we found that the null hypothesis of the first hypothesis test was rejected. Therefore, in order to find the which supermarket was cheaper, we performed the 2nd hypothesis test.

The null hypothesis (H0) of the second hypothesis test is the assumption that the product prices of both stores are equal. The alternative hypothesis (HA) of the second hypothesis test is the assumption that the product price of Coles is cheaper than the product price of Woolworths.

Hide

```
#1st Paired T-test to perform hypothesis test
t.test(Store$ColesPrice, Store$WoolworthsPrice,
       paired = TRUE,
       alternative = "two.sided")
```

```
    Paired t-test

data:  Store$ColesPrice and Store$WoolworthsPrice
t = -2.0121, df = 79, p-value = 0.04762
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.670624804 -0.003625196
sample estimates:
mean of the differences
              -0.337125
```

Hide

```
#2nd Paired T-test to find if coles is cheaper than woolworths
t.test(Store$ColesPrice, Store$WoolworthsPrice,
       paired = TRUE,
       alternative = "less")
```

```
    Paired t-test

data:  Store$ColesPrice and Store$WoolworthsPrice
t = -2.0121, df = 79, p-value = 0.02381
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        -Inf -0.05825968
sample estimates:
mean of the differences
              -0.337125
```

Hide

```
#checking the critical value
qt(p = 0.025, df = 80)
```

```
[1] -1.990063
```

Hide

```
#Checking the p-value
2*pt(q = -1.99, df = 80)
```

```
[1] 0.05000713
```

Interpretation

A paired t-test was used to test for a significant difference between the mean product prices of Coles and Woolworths. Since the mean difference was found to be -0.34 (SD = 1.49), we reject the null hypotheses. Visual inspection of the Q-Q plot suggested that the data were not normally distributed but due to the large dataset, we were fortunate enough to ignore the non-normality present in the data.The paired-samples t-test

found a statistically significant mean difference between the price of two stores, t(df=80) = –2.012, p < 0.048, 95% [-0.671 , - 0.004] . Using the p-value approach, since p-value = 0.04762 was less than 0.05 significance level, we reject the null hypothesis that product prices of both stores are equal.

A second paired t-test was used to test for a significant difference between the mean product prices of Coles and Woolworths.The second paired-samples t-test found a statistically significant mean difference between the price of two stores, t(df = 80) = –2.012, p < 0.024, 95% [-INF , -0.059] . Using the p-value approach, since p-value = 0.024 was less than 0.05 significance level, we reject the null hypothesis that product prices of both stores are equal. Since the null hypothesis was rejected, the results of the hypothesis test have provided statistically significant evidence to support the assumption that the product price of Coles supermarket is cheaper than the Woolworths.

Discussion

The strategic analysis of the data collected, clearly showed that approximately, half the majority of products were found cheaper at Coles over Woolworths. We would like to clearly point out the constraint that our analysis was confined to the random samples of product price we possessed. The stratified sampling approach for random sampling by dissecting the products into 7 such sub-categories namely, Fruits, Breakfast, Frozen, Vegetable, Snacks, Dairy, and Raw. Extensive analysis of data, keeping both Coles prices and Woolworths prices with comparable weight to each subcategory, we investigated to find the inclination to the side where the price is lower than the other for the same category of product. The strength of our analysis lies with the categories we choose for analysis as they are most readily available and favored products which can be found in almost any supermarket. Thus, making the scope for expanding the analysis to other supermarkets too if required. The major limitation was the span of products our data set covered, which is just a minute share compared to the number of products these two supermarket giants has to offer. Finally, the major improvement we would like to make is feeding in more categories and increasing the number of products to provide a bigger pool to execute on.