MATH1324 Assignment 2

Code

Supermarket Price Wars

Group/Individual Details

- Ashwin Anis (s3763476)
- Deepak Prasad (s3759108)

Executive Statement

This report contains a statistical analysis of the prices between Coles and Woolworths, two of Australia's leading supermarket chains. This study is to assess which supermarket chain is cheaper than the other based on the mean of the prices of products available at both the supermarkets.

The data required for the study was collected online from the respective websites of each supermarket. The dataset contains three attributes namely product name, price, and company. The dataset was then split by grouping the observations according to their respective supermarkets. Summary statistics have been presented as part of descriptive statistics and a two-sample t-test for independent variables was carried out for statistical inference.

In descriptive statistics, mean, median, mode,1st quartile, and other features have been explicated and box plot is used to visualize the same. The normality of the data is visualized using q-q plots and we could see that the majority of the observations lay inside the confidence interval. However, we considered CLT to assume normality of the distribution as the sample size was 400(n>30).

The Levene's test was applied to the sample to check the homogeneity of the variances, and the resulting p-value was 0.4064 which is greater than 0.05, so we are safe to assume that variances are equal. Following this, the two sample t-test was carried out and the resulting p-value (0.60) was found to be greater than 0.05 meaning that we failed to reject the null hypothesis, thereby getting to the conclusion that there is no statistically significant difference in the means of the prices of both supermarkets to suggest one is cheaper than the other and requires more evidence to prove otherwise.

Load Packages and Data

Hide

```
#Importing Libraries
library(rmarkdown)
library(readxl)
library(magrittr)
library(dplyr)
library(car)
library(kableExtra)
#Reading Excel File
data <- read_excel("/Users/dprasadg/Downloads/Statistics2.xlsx")
```

```
-
/
```

Hide

```
View(data)
data$Company <- factor(data$Company, levels= c("Coles","Woolys"), labels=c("c","w"))
```

Summary Statistics

Use R to summarise the data from the investigation. Include an appropriate plot to help visualise the data. Describe the trend.

Hide

```
a <- data %>% group_by(Company) %>%
  summarise(Observations = n()
            ,Mean = mean(Price)
            ,Median= median(Price)
            ,'Std. Deviation' = sd(Price)
            ,'1st Quantile' = quantile(Price, .25)
            ,'3rd Quartile' = quantile(Price, .75)
            ,'Inter Quartile' = quantile(Price, .75) - quantile(Price, .25)
            ,Minimum = min(Price)
            ,Maximum = max(Price)
            ,Missing = sum(is.na(Price)))
kable(a, caption = "Descriptive statistics on prices based on companies") %>%
  kable_styling(bootstrap_options = c("hover", "condensed"))
```

Descriptive statistics on prices based on companies

Company	Observations	Mean	Median	Std. Deviation	1st Quantile	3rd Quartile	Inter Quartile	Minimum	Maximum	Missing
c	200	4.32740	4	2.092175	2.7000	5.8	3.1000	1.0	10	0
w	200	4.41515	4	2.174151	2.9975	5.7	2.7025	0.8	10	0

Hide

```
data %>% boxplot(Price ~ Company, data = ., ylab = "Price",xlab = "Companies", main = "Box pl
ot for comparision of prices between coles and woolworths")
```

Hypothesis Test

Use R to perform an appropriate hypothesis test to determine which supermarket is the cheapest. You need to explain your choice of hypothesis test, any assumptions and the significance level.

Hide

```
#Sample
set.seed(1)
samp <- sample_n(data,size=400)
samp
```

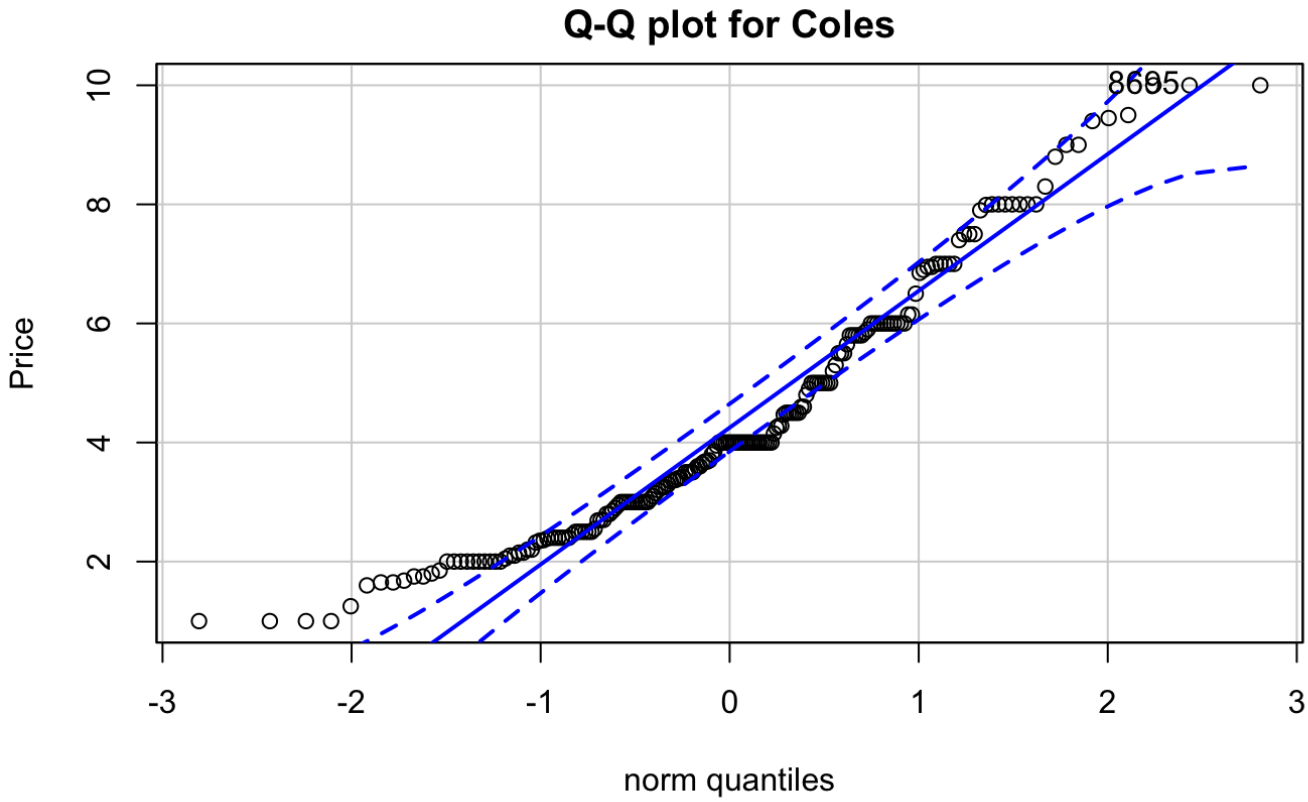
Product Name <chr>	Price <dbl>	Comp… <fctr>
Kellogg's Just right	4.00	c
Maybelline Eye & Lip Makeup Remover	6.00	c
Kellogg's All Bran Original	4.29	w
Sustagen Dutch Chocolate Drink Powder	7.50	w
ORGANIC CARE Anti Dandruff 2 in 1 Shampoo & Conditioner	3.50	c
Ambi Pur Air Effects Lavender Vanilla & Comfort Air Freshener	6.00	w
Ovaltine Drinking Chocolate	7.00	w
Kraft Easy Mac Cheese Microwavable Macaroni Snack Multi Pack	4.00	w
Colgate 360 Degree Soft Toothbrush	5.00	w
Continental Bacon Carbonara Pasta & Sauce	2.10	c

1-10 of 400 rows Previous **1** 2 3 4 5 6 … 40 Next

Hide

```
#Q-Q plot for Coles
samp_coles <- samp %>% filter(Company == "c")
samp_coles$Price %>% qqPlot(dist="norm", main = "Q-Q plot for Coles", ylab = "Price")
```

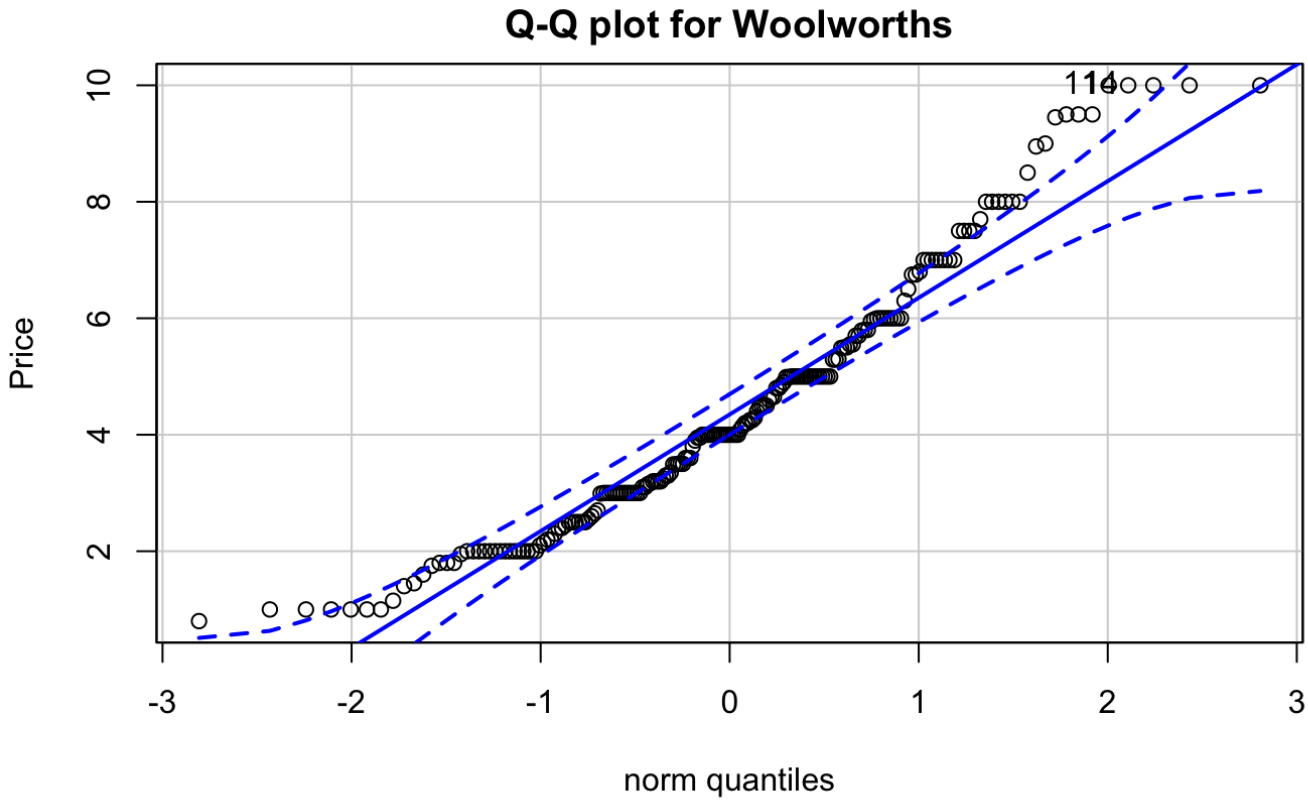
```
[1] 86 95
```



Hide

```
#Q-Q plot for Woolworths
samp_w <- samp %>% filter(Company == "w")
samp_w$Price %>% qqPlot(dist="norm", main = "Q-Q plot for Woolworths", , ylab = "Price")
```

```
[1] 11 14
```



Hide

```
#Levene's Test to check homogeneity of variance
leveneTest(Price ~ Company, data = samp)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.3588 0.5495
      398
```

Hide

```
#Indepentent two sample t-test for independent variables with equal variance
t.test(
  Price ~ Company,
  data = samp,
  var.equal = TRUE,
  alternative = "two.sided"
)
```

```

    Two Sample t-test

data:  Price by Company
t = -0.41129, df = 398, p-value = 0.6811
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.507195  0.331695
sample estimates:
mean in group c mean in group w
        4.32740         4.41515
```

Interpretation

An Independent two-sample t-test was used to test for a significant difference between the mean prices of Coles and Woolworths. While the prices for Woolworths exhibited evidence of non-normality upon inspection of the normal Q-Q plot, the central limit theorem ensured that the t-test could be applied due to the large sample size in each group. The Levene's test of homogeneity of variance indicated that equal variance could be assumed. The results of the two-sample t-test assuming equal variance found the difference between the mean prices of Coles and Woolworths are not statistically significant, t(df=398)=-0.4113, p=0.68, 95% CI for the difference in means [-0.51 0.33]. The results of the investigation suggest that to find which of the two supermarkets, Woolworths or Coles is cheaper, requires more evidence to deduce the statistically significant results.

Discussion

After completing this analysis, we came to the conclusion that there is no statistically significant difference in the mean of both the supermarkets to suggest which is cheaper than the other.

One of the strengths of this anlaysis was that we made sure to include a diverse range products in the dataset so that the data is not baised to a particular category of products.We also introduced an upper price limit of 10$ so that there are no huge variations between the prices of the products in the dataset.

As far as the weaknesses are concerned, we feel that the sample size could have been bigger thereby increasing the possibility of the null hypothesis being rejected due to the decrease in the p-value. Another weakness was that there were a fair number of outliers visible in the Q-Q plots of both supermarkets.