MATH1324 Assignment 2

Code

Supermarket Price Wars

Group Details

- Subhasree Mohapatra (s3779215)
- Rutvi Macwan (s3773570)

Executive Statement

The aim of the below investigation is to compare the pricing of various products between the two Supermarkets, COles and Woolsworth and to conclude which has the cheaper price rates for products in the market. To analyse the pricing,a total of 76 products has been considered as sample size from respective supermarkets(152 prices has been noted in total).We intend to take a large sample for the reason to minimize the standard error of the population as a large sample approaches to have a normal distribution.The selection has been made on the basis of availabilty of products in the supermarkets by considering the trends on websites of the supermarkets: www.coles.com.au and www.woolworths.com.au. The selection has been done randomly from various categories of available products in both the supermarket which includes various category such as variou Food categories,Baby Products,Beauty Products.Toiletries .Due to the unavailibity of many products in one of the super market or variation in sales strategy for brands hence ,we were not able to have an full list of product list of the supermarkets hence we chose about the items which were readily available in both the market to ensure the equivalence in the quality,quantity,ingredients and other properties.The methods that are being used for summary statistics will be descriptive statistic and the box-plotting,we will be also be demonstrating whether the prices are normally distributed or not by various methods mainly the Q-Q Plot,Shapiro-Wilk test,Histogram and provide the assumption for the distribution .We will be performing the paired sample t-test for the pricing investigation with giving valid reason of choosing this test for the investigation during the test below. ## Load Packages and Data

Hide

```
# This is a chunk where you can load the necessary data and packages required to reproduce the r
eport
# You should also include your code required to prepare your data for analysis.
library(dplyr)
library(ggplot2)
library(magrittr)
library(car)
library(tidyr)
library(outliers)
Sprices<-read.csv("C:/Users/DELL/Documents/SuperMarketPrices.csv")
ColesPrice<-Sprices$Coles
WooliesPrice<-Sprices$Woolworth
Supermarket_com<-mutate(Sprices,PriceDiff=ColesPrice-WooliesPrice)
head(Supermarket_com)
```

Products	Coles	Woolworth	PriceDiff
<fctr>	<dbl>	<dbl>	<dbl>

Products <fctr>	Coles <dbl>	Woolworth <dbl>	PriceDiff <dbl>
1 Pineapple	3.20	3.9	-0.70
2 Mandarin	7.14	6.5	0.64
3 South Cape Greek Feta	5.00	5.0	0.00
4 Watermelon	4.25	4.7	-0.45
5 Steggles sweet chilli chicken tender	11.00	11.0	0.00
6 Frozen Crumbed Chicken Breast Fingers	7.00	9.5	-2.50

6 rows

Code

Summary Statistics

The below code and technique summarises the pricing of Woolworth and Coles . We can see that the mean price value of Coles is 5.62 and the mean price value of Woolsworth is 5.61 approximately . Studying the other descriptive statistic variables and the box plot , we can see there are very less difference between the prices . The prices for both Coles and Woolsworth tends be much close. In the box-plot we can visualise that there is a less difference between the prices but we can clearly see one outlier in the prices of coles which we will be removing by imputing process to further assess the out come . In the Q-Q plotting performed it can be seen there are some values which are deviating from the normality in both pricing but since we have taken account a large dataset sample (n=76) and we can visualize that maximum datapoints falls close to normal distribution range(between the dashed lines) which indicates that they are in 95% CI interval and some point are placed on the dashed lines .. Similarly in the shapiro test, the p value of is small which indicates deviation from normal distribution but we can consider the sample for further analysis on the assumption of Central Limit Theorem. From the histogram study with mean of sample marked by the blue solid line on the graph, we can see that the pricing distribution are not normal(bell-shaped curve) but are tending to approach normality. In addition to the above,as per the Central Limit Theorem, when the sample is large(n>30) it can be assumed that the sampling distribution will be having a approximate mean close to population mean regardless of the underlying population distribution.Since we have a large dataset of sample size =76, we can work on the hypothesis testing according to the Central Limit Theorem and consider the sample for t-test In the Line plot, we have shown the trend of prices of both the supermarkets compared to each other on basis of products..

Hide

```
Supermarket_com%>%summarise(Mean=mean(Coles,na.rm=TRUE),
                    Median=median(Coles,na.rm=TRUE),
                    Q1=quantile(Coles,prob=0.25,na.rm=TRUE),
                    Q3=quantile(Coles,prob=0.75,na.rm=TRUE),
                    MaxVal=max(Coles,na.rm=TRUE),
                    MinVal=min(Coles,na.rm=TRUE),
                    IQR=Q3-Q1,
                    SD=sd(Coles,na.rm=TRUE),
                    Range=MaxVal-MinVal)
```

Mean <dbl>	Median <dbl>	Q1 <dbl>	Q3 <dbl>	MaxVal <dbl>	MinVal <dbl>	IQR <dbl>	SD <dbl>	Range <dbl>
5.616316	5	3.5	7.23	15	1	3.73	2.752042	14

1 row

Hide

```
Supermarket_com%>%summarise(Mean=mean(Woolworth,na.rm=TRUE),
                    Median=median(Woolworth,na.rm=TRUE),
                    Q1=quantile(Woolworth,prob=0.25,na.rm=TRUE),
                    Q3=quantile(Woolworth,prob=0.75,na.rm=TRUE),
                     SD=sd(Woolworth,na.rm=TRUE),
                    MaxVal=max(Woolworth,na.rm=TRUE),
                    MinVal=min(Woolworth,na.rm=TRUE),
                    IQR=Q3-Q1,
                    Range=MaxVal-MinVal)
```

Mean <dbl>	Median <dbl>	Q1 <dbl>	Q3 <dbl>	SD <dbl>	MaxVal <dbl>	MinVal <dbl>	IQR <dbl>	Range <dbl>
5.606053	5	3.765	7.5	2.614452	12	1.5	3.735	10.5

1 row

Hide

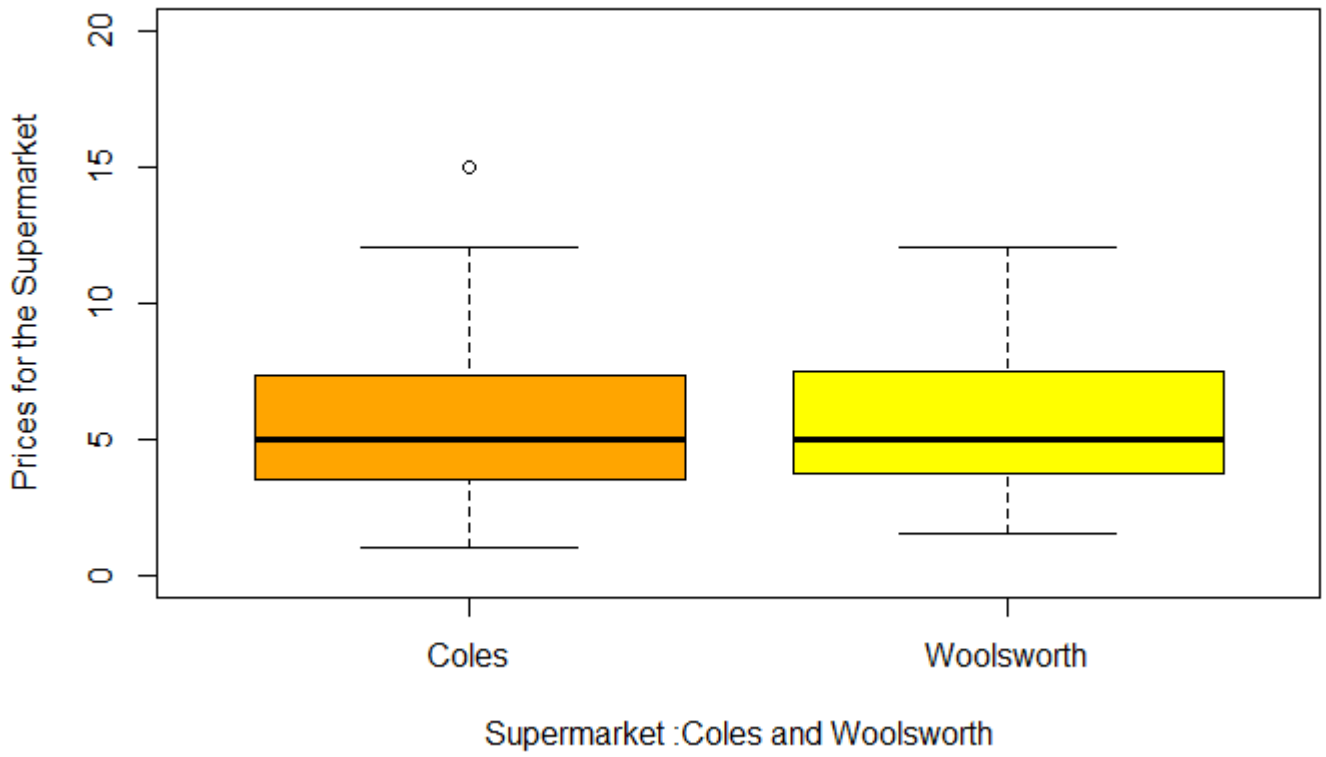
```
Supermarket_com%>%summarise(Mean=mean(PriceDiff,na.rm=TRUE),
                    Median=median(PriceDiff,na.rm=TRUE),
                    Q1=quantile(PriceDiff,prob=0.25,na.rm=TRUE),
                    Q3=quantile(PriceDiff,prob=0.75,na.rm=TRUE),
                     SD=sd(PriceDiff,na.rm=TRUE),
                    MaxVal=max(PriceDiff,na.rm=TRUE),
                    MinVal=min(PriceDiff,na.rm=TRUE),
                    IQR=Q3-Q1,
                    Range=MaxVal-MinVal)
```

Mean <dbl>	Median <dbl>	Q1 <dbl>	Q3 <dbl>	SD <dbl>	MaxVal <dbl>	MinVal <dbl>	IQR <dbl>	Range <dbl>
0.01026316	0	-0.0225	0	1.175783	5	-3.6	0.0225	8.6

1 row

Hide

```
boxplot(Supermarket_com$Coles,Supermarket_com$Woolworth,names=c("Coles","Woolsworth"),xlab="Supe
rmarket :Coles and Woolsworth",ylab="Prices for the Supermarket",col=c("orange","yellow"),ylim=c
(0,20))
```



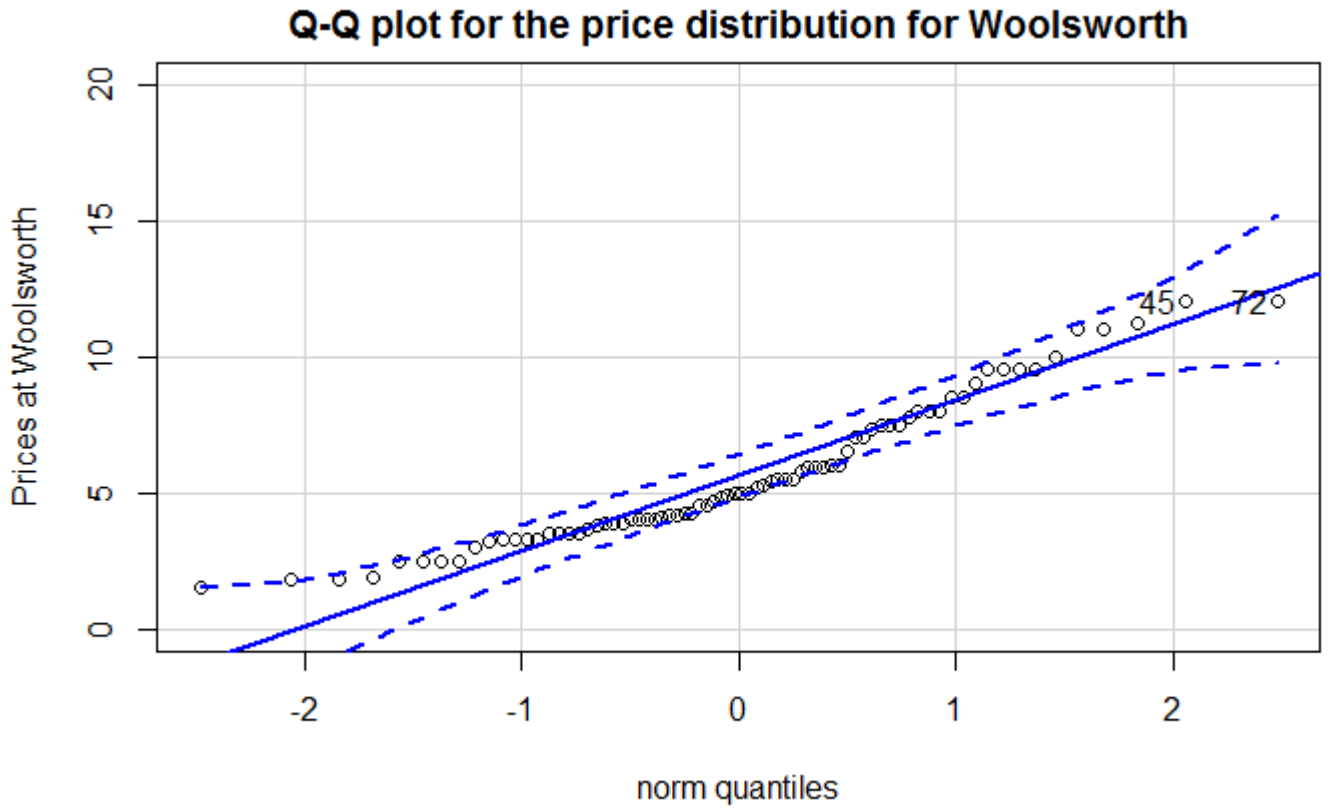
Hide

```
z.scores <- Supermarket_com$Coles %>%  scores(type = "z")
Supermarket_com$Coles[ which( abs(z.scores) >3 )] <- mean(Supermarket_com$Coles, na.rm = TRUE)

boxplot(Supermarket_com$Coles,Supermarket_com$Woolworth,names=c("Coles","Woolsworth"),xlab="Supe
rmarket :Coles and Woolsworth after removing outlier",ylab="Prices for the Supermarket",col=c("o
range","yellow"),ylim=c(0,20))
```

Hide

```
#normality test
qqPlot(Supermarket_com$Woolworth,dist="norm",main="Q-Q plot for the price distribution for Wools
worth",ylab="Prices at Woolsworth",ylim=c(0,20))
```

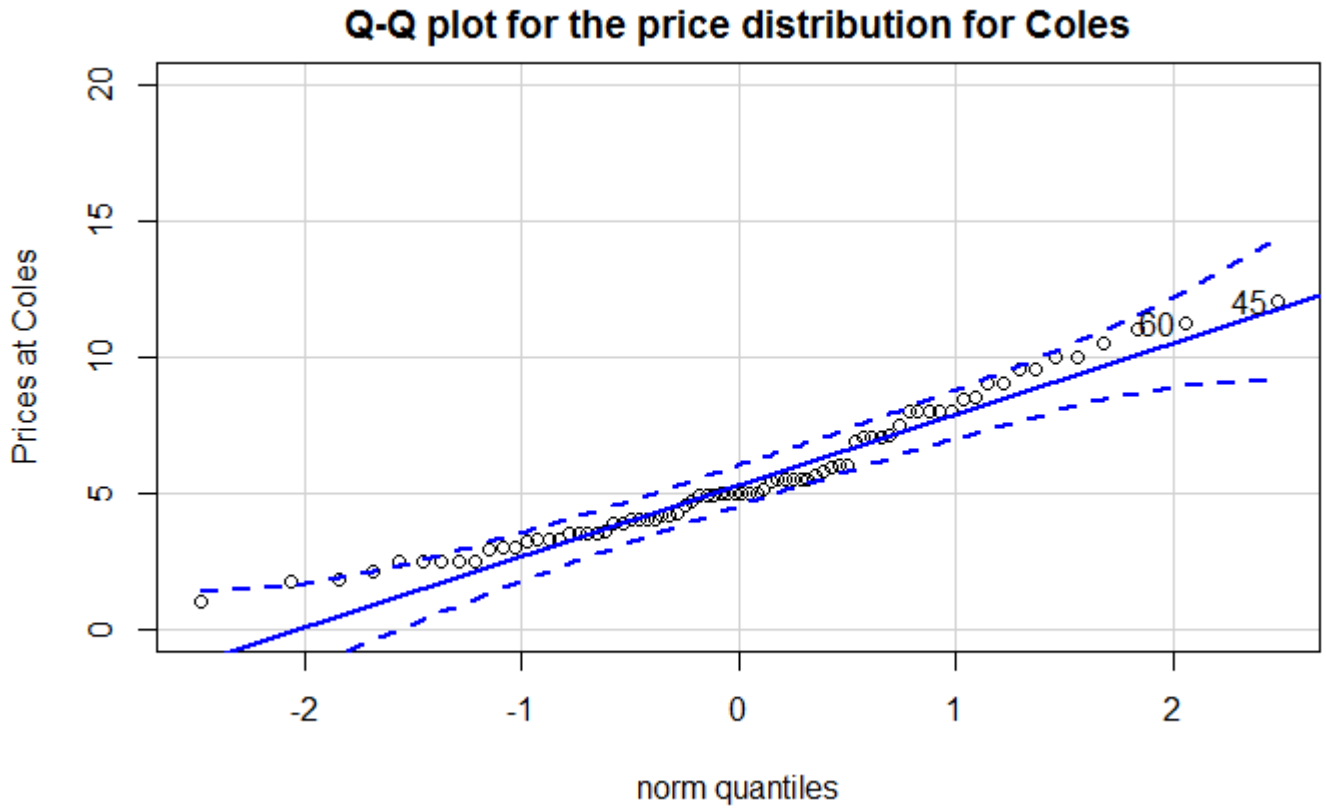
```
[1] 45 72
```



Hide

```
qqPlot(Supermarket_com$Coles,dist="norm",main="Q-Q plot for the price distribution for Coles",yl
ab="Prices at Coles",ylim=c(0,20))
```

```
[1] 45 60
```



Hide

```
shapiro.test(Supermarket_com$Coles)
```

```
    Shapiro-Wilk normality test

data:  Supermarket_com$Coles
W = 0.951, p-value = 0.005259
```

Hide

```
shapiro.test(Supermarket_com$Woolworth)
```

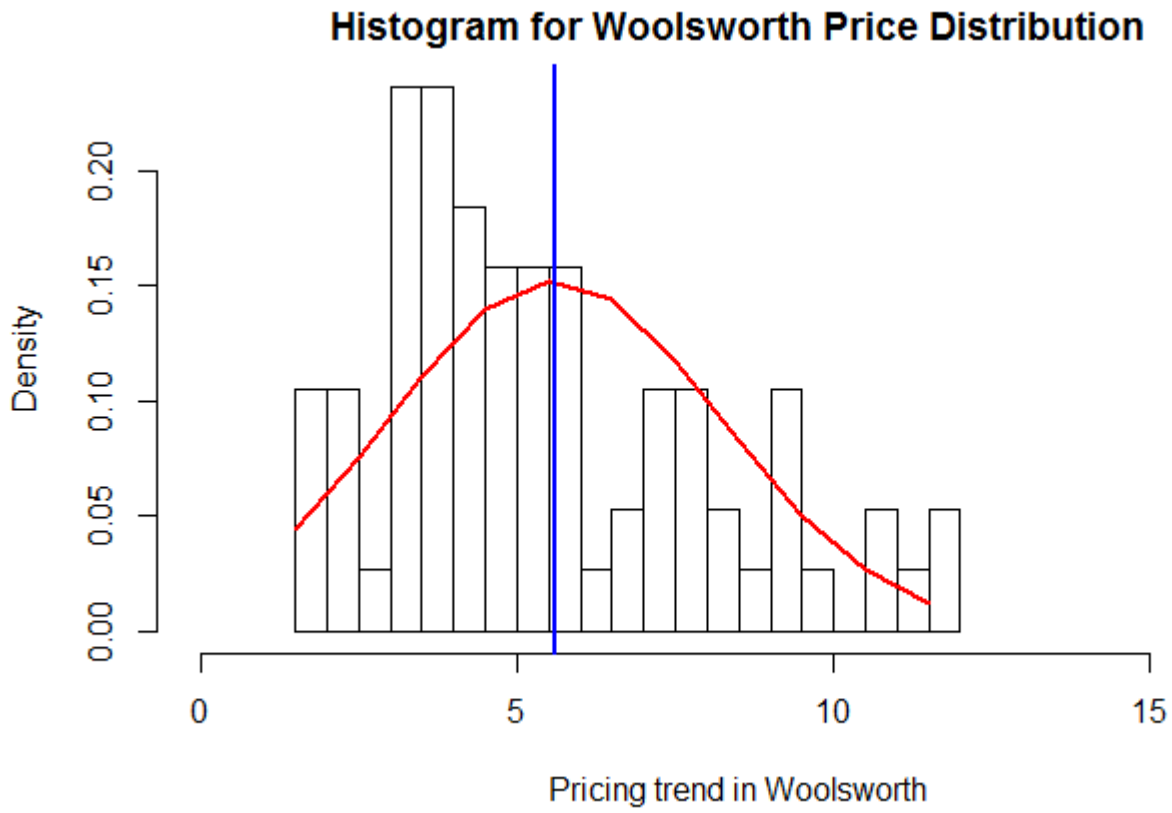
```
    Shapiro-Wilk normality test

data:  Supermarket_com$Woolworth
W = 0.9359, p-value = 0.0008299
```

Hide

```
Wmean=mean(Supermarket_com$Woolworth,na.rm = TRUE)
Wmed=median(Supermarket_com$Woolworth,na.rm = TRUE)
WMax=max(Supermarket_com$Woolworth,na.rm = TRUE)
WMin=min(Supermarket_com$Woolworth,na.rm = TRUE)
Wsd=sd(Supermarket_com$Woolworth,na.rm = TRUE)
Wdorm=dnorm(WMin:WMax,mean = Wmean,sd=Wsd)
hist(Supermarket_com$Woolworth,probability=TRUE,xlim = c(0,17),breaks=30,xlab="Pricing trend in
 Woolsworth",main ="Histogram for Woolsworth Price Distribution" )
lines(x=WMin:WMax,y=Wdorm,lwd=2,col="red")
```

Hide

```
abline(v = Wmean,col = "blue",lwd = 2)
```

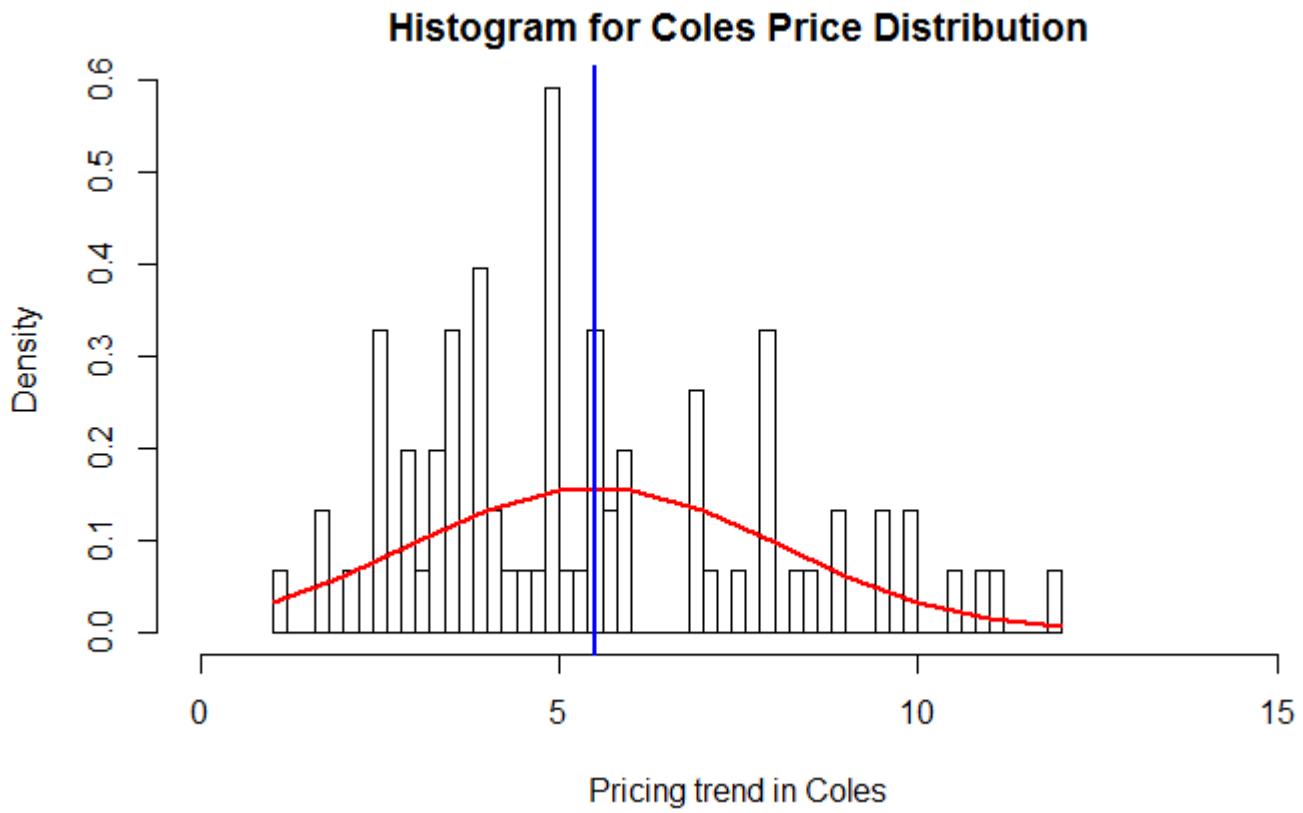


Hide

```
Cmean=mean(Supermarket_com$Coles,na.rm = TRUE)
Cmed=median(Supermarket_com$Coles,na.rm = TRUE)
CMax=max(Supermarket_com$Coles,na.rm = TRUE)
CMin=min(Supermarket_com$Coles,na.rm = TRUE)
Csd=sd(Supermarket_com$Coles,na.rm = TRUE)
Cdorm=dnorm(CMin:CMax,mean = Cmean,sd=Csd)
hist(Supermarket_com$Coles,probability=TRUE,xlim = c(0,15),breaks=45,xlab="Pricing trend in Cole
s",main ="Histogram for Coles Price Distribution" )
lines(x=CMin:CMax,y=Cdorm,lwd=2,col="red")
```

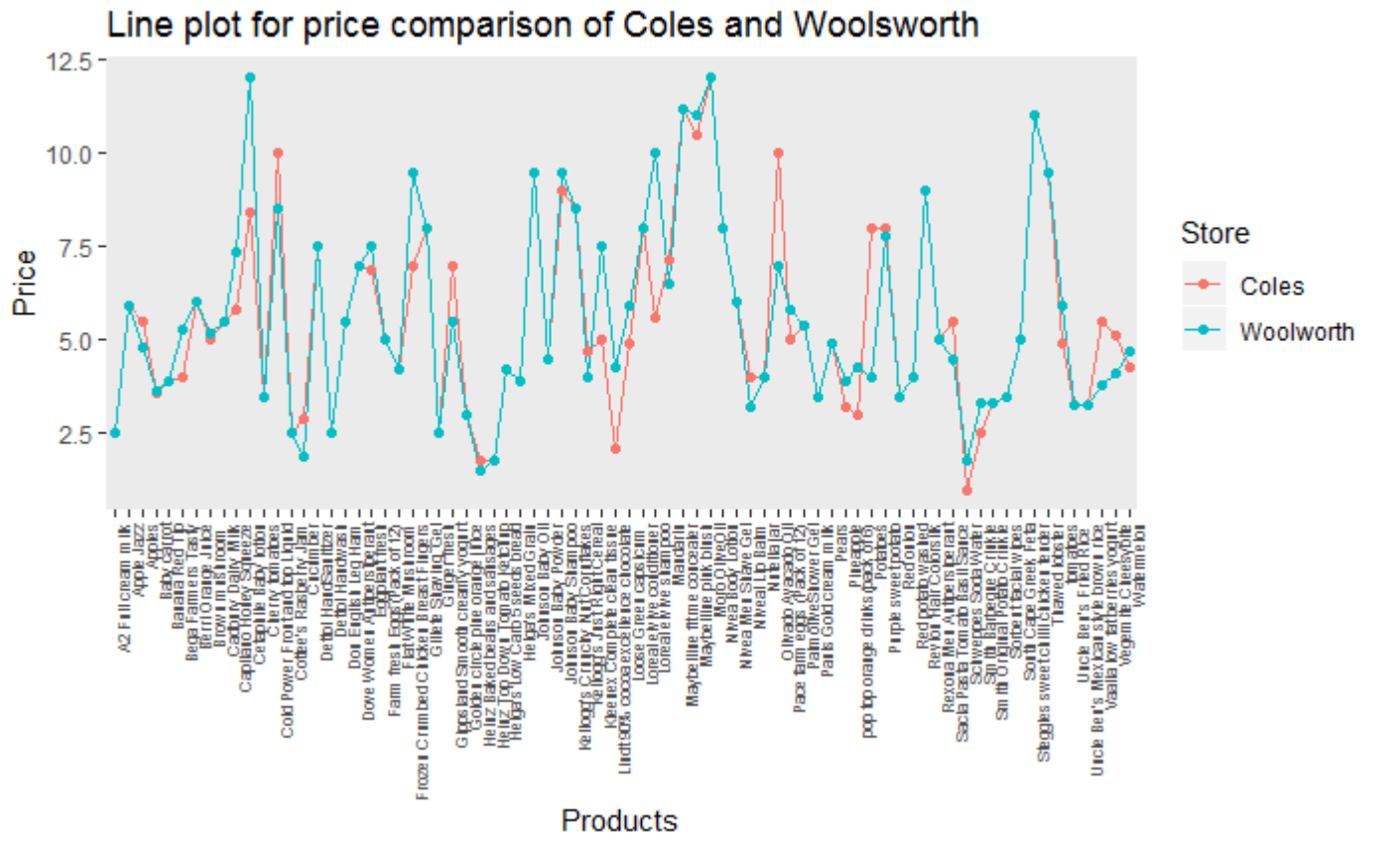
Hide

```
abline(v = Cmean,col = "blue",lwd = 2)
```



Hide

```
price_compare<-Supermarket_com%>%gather(Coles,Woolworth,key="Store",value="Price",convert=TRUE,na.rm=TRUE)
ggplot(data=price_compare,
       aes(x=Products,y=Price,group=Store,colour=Store))+
        geom_line()+geom_point()+ggtitle("Line plot for price comparison of Coles and Woolsworth")+theme(axis.text.x=element_text(size=6,angle = 90,vjust=1,hjust=1),

panel.grid.major.x = element_blank(),panel.grid.minor.x = element_blank(),panel.grid.major.y = element_blank(),panel.grid.minor.y = element_blank() )
```



Code

Hypothesis Test

The hypothesis testing for the given investigation is paired t-test because the selection of the samples for pricing were not independent as we were focussing on selection of products to ensure equivalency between the same products from both the supermarket,hence we are sampling the price of similar product having same brand,quantity,quality etc.We will be assuming that for a large sample size for t-distribution the sample tends to normality with mean =0,and standard deviation =1. We will be assuming the significance level α=0.05 and the 95% CI interval.We will be taking the Null hypothesis as the mean differences between the pricing to be equal to 0 and the alternative hypothesis as mean difference between the pricing to be not equal to 0.We will taking two-tailed test for the investigation.

According to the paired t-test:

The mean is Δμ

Ho(Null Hypothesis):Δμ=0(Pricing mean diffrence for Coles and Woolsworth is 0)

Ha(Alternate Hypothesis):Δμ≠0

Hide

```
pairedttest <- t.test(Supermarket_com$Coles,Supermarket_com$Woolworth, paired=TRUE,
       conf.level = 0.95, alternative = "two.sided")
pairedttest
```

```
    Paired t-test

data:  Supermarket_com$Coles and Supermarket_com$Woolworth
t = -0.86809, df = 75, p-value = 0.3881
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3729926  0.1465798
sample estimates:
mean of the differences
             -0.1132064
```

Hide

```
#crtical t

qt(p=0.025,df=75)
```

```
[1] -1.992102
```

Hide

```
qt(p=(1-0.025),df=75)
```

```
[1] 1.992102
```

Interpretation

By performing the paired-samples t-test for significant difference in prices mean for both Coles and Woolworths. The mean difference between two stores was found to be -0.113(approx) with the standard deviation of 1.176(approx). The p-value from t-test turns out to be 0.388(approx) which greater than the significant level α=0.05, and the confidence interval[-0.373,0.147] that contains the null hypothesis value of 0.We find that the t statistics to be -0.868 and t-criticals are lower critical point at -1.992 and upper critical point at 1.992 beyond which the null hypothesis will be falling in the rejection area but our t-statistic does not fall in these area. Thus, comparison of prices was found to be not statistically significant or it fails to reject null hypothesis, its hard to say which supermarket bears a cheaper rate. This requires further investigation.

Discussion

Although , we have taken a large dataset but there some limitations for which it cannot be completely said that it is a representative sample because we do not have a complete set of product pricing of the supermarkets because the selective product marketing based on various brands or different category,which are dissimilar in both the super markets in many places,many products were out of stock and not replenished,our research being limited to some online available products under some catergories which were available in both the supermarkets, where as there may be much larger sales in the supermarket stores.The products offered varies from regions to regions.Hence to have a sample approximatelyrepresentative of the pricing we need to accumulate all the pricing of all categories of products overall.