MATH1324 Assignment 2

Supermarket Price Wars

Group/Individual Details

- Aasrita Kalahasti(s3800189)
- Vignesh Gopalakrishnan(s3795594)
- Kanishka Tamang (s33756188)

Executive Statement

Purpose of Investigation:

- The objective of the report is to understand if prices of Coles and Woolworths are statistically different and to try arriving at a conclusion as to which one is cheaper.
- The data used for analysis has been collected from http://www.grocerycop.com.au/products (http://www.grocerycop.com.au/products), it's a website that compares different product's prices in Coles and Woolworths. Hence, choosing this website ensures that the products are matched.
- Actual price of each product has been taken into account because both stores don't run discounts of same products at the same given point of time.

Sample Description:

- The sample has 269 observations consiting of 4 variables: category, product, Coles and Woolworths.
- The data is spread across 9 categories.
- Data collection has been done by each group member and collecting about 30 observations from each category has ensured randomness.
- The products chosen are exactly the same in terms of quality, quantity and brand.
- An adequate sample size is different in one's perspective, collecting atleast 30 observations from each category in order to reduce standard error seemed adequate for our group and hence, 269.

Method Of Investigation:

- Grouped Bar Chart is being plotted to get a clear understanding of prices of products when grouped by categories.
- Histogram with normal distribution overlay and Boxplot will be plotted to clean the data and get a visual understanding of differences in prices.
- Q-Q Plot will help us in checking the normality of the data.

- The paired Sample T-Test is the ideal test in this case to find out if the average prices of the products of Coles and Woolworths is same.

- This particular test is chosen for the investigation on a rationale that the price of a product at coles will be inherently linked to the price of the same product at woolworths.

Load Packages and Data

```
library(readxl) #To read the data
library(dplyr) #Data manipulation
```

```
##
## 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(knitr) #To create dynamic reports
library(ggplot2) #To view complex plots
library(tidyr)#To tidy the data
library(car) #qqplot
```

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

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

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

```
library(magrittr) #To use pipe operator
```

```
##
## Attaching package: 'magrittr'
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
#Reading the Data
prices_data <- read_xlsx("Groceriesdata.xlsx")
dim(prices_data)
```

```
## [1] 269   4
```

```
#Preparing data for analysis
prices_data <- prices_data %>% mutate(Difference = Coles - Woolworths)
prices_data
```

Category <chr> ▸
Baby,Health& Beauty
Baby,Health& Beauty
Baby,Health& Beauty
Baby,Health& Beauty
Baby,Health& Beauty
Baby,Health& Beauty
Baby,Health& Beauty
Baby,Health& Beauty
Baby,Health& Beauty
Baby,Health& Beauty

1-10 of 269 rows | 1-1 of 5 columns Previous **1** 2 3 4 5 6 ... 27 Next

```
gathered_supermarts <- prices_data %>%
gather("Coles","Woolworths",key = "Store", value = "price")
gathered_supermarts$Store <- as.factor(gathered_supermarts$Store)
```

Summary Statistics

- The mean Coles price is found out to be 10.561 and that for Woolworths is 10.312,therefore the mean difference(Coles - Woolworths) is 0.249 and the standard deviation is 2.417 which gave us an indication that Woolworths is cheaper than Coles.

- From the bar chart,we understood the prices of Coles and Woolworths vary with categories.
- In three categories,Woolworths is more expensive than Coles: Meat and Seafood, Entertainment and International, Drinks and Tobacco.
- In all the other categories,Coles is found to be expensive.
- When plotted a histogram with a normal distribution overlay of differences of the prices it is found that the graph is right skewed.
- A Category wise Boxplot has been plotted and it clears shows that price of Drinks and Tobacco is much higher in Coles than that of Woolworths, price of the categories Fridge, Bakery and Entertainment and International is almost the same in both stores.
- As per the Q-Q plot, it’s been observed that the data isn’t normally distributed.
- However,collection of 269 observations which is greater than 30 has given a scope to assume normality as per Central Limit Theorem.
- In order to remove the outliers,boxplot of differences of prices has been plotted.
- The upper fence is calculated to be 1.28 and lower fence to be -0.8.
- All the values above and below the fences have been removed which lead towards getting a tidy data without the outliers to perform the hypothesis testing.

```
# Summary of prices (Supermarkets)

summary_table <- gathered_supermarts %>% group_by(Store) %>%
summarise(Min = min(price,na.rm = TRUE) %>% round(2),
            Q1 = quantile(price,probs = .25,na.rm = TRUE) %>% round(2),
            Median = median(price, na.rm = TRUE) %>% round(2),
            Q3 = quantile(price,probs = .75,na.rm = TRUE) %>% round(2),
            Max = max(price,na.rm = TRUE) %>% round(2),
            Mean = mean(price, na.rm = TRUE) %>% round(3),
            SD = sd(price, na.rm = TRUE) %>% round(3),
            n = n())
summary_table
```

Store <fctr>	Min <dbl>	Q1 <dbl>	Median <dbl>	Q3 <dbl>	Max <dbl>	Mean <dbl>	SD <dbl>	n <int>
Coles	1	4	6.0	9.99	123	10.561	15.540	269
Woolworths	1	4	5.9	9.30	139	10.312	16.376	269

2 rows

```
## Summary of difference in prices (Supermarkets)

price_difference_summary <- prices_data %>%
summarise(Min = min(Difference,na.rm = TRUE) %>%
          round(2),Q1 = quantile(Difference,probs = .25,na.rm = TRUE) %>%
          round(2),Median = median(Difference, na.rm = TRUE) %>%
          round(2),Q3 = quantile(Difference,probs = .75,na.rm = TRUE) %>%
          round(2),Max = max(Difference,na.rm = TRUE) %>%
          round(2),Mean = mean(Difference, na.rm = TRUE) %>%
          round(3),SD = sd(Difference, na.rm = TRUE) %>%
          round(3),n = n())
price_difference_summary
```

Min <dbl>	Q1 <dbl>	Median <dbl>	Q3 <dbl>	Max <dbl>	Mean <dbl>	SD <dbl>	n <int>
-18	-0.02	0	0.5	9.54	0.249	2.417	269

1 row

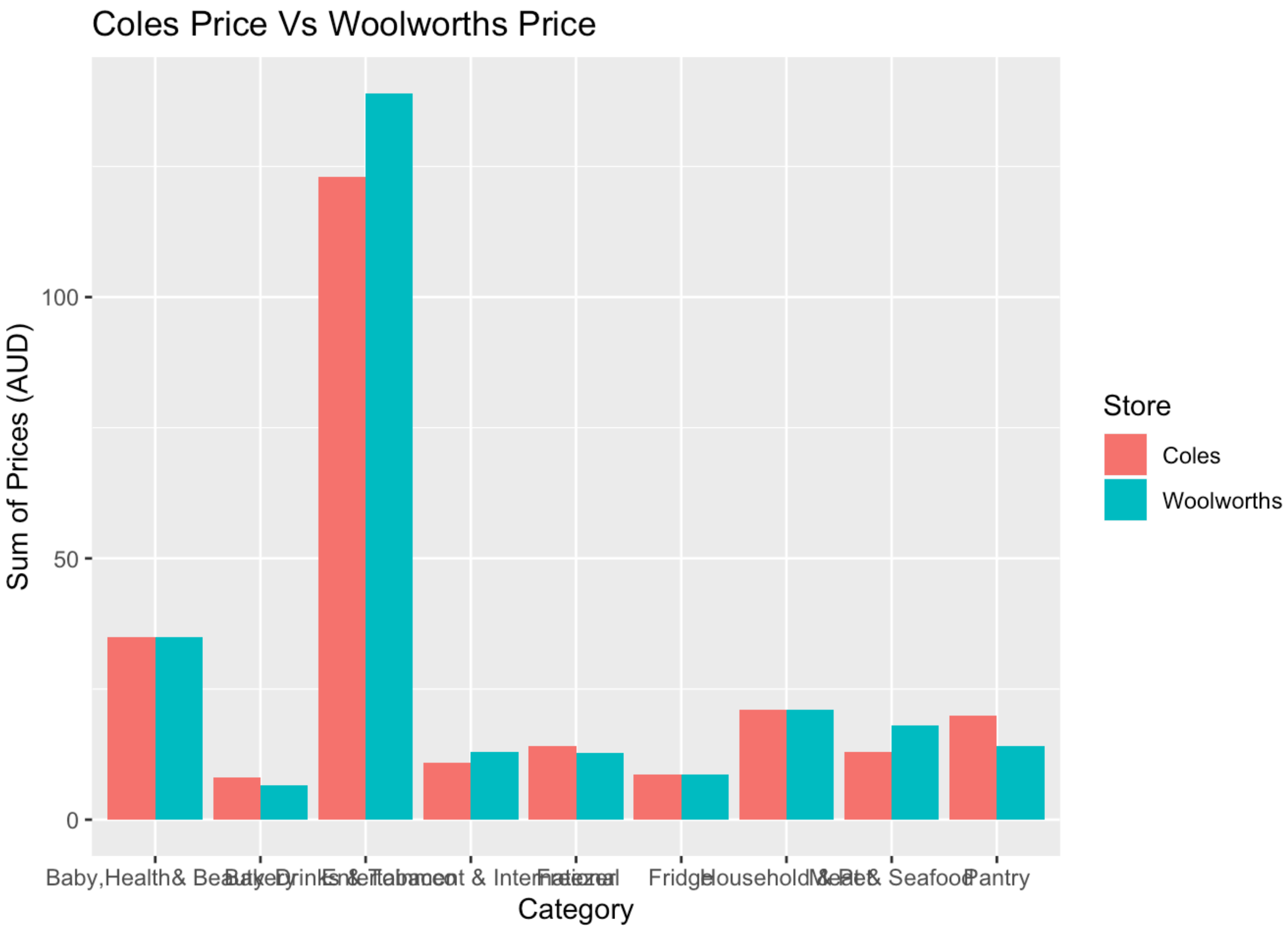
```
#Summary of Difference grouped by Category

price_difference_summary_category <- prices_data %>%
group_by(Category) %>%
summarise(Min = min(Difference,na.rm = TRUE) %>%
          round(2),Q1 = quantile(Difference,probs = .25,na.rm = TRUE) %>%
          round(2),Median = median(Difference, na.rm = TRUE) %>%
          round(2),Q3 = quantile(Difference,probs = .75,na.rm = TRUE) %>%
          round(2),Max = max(Difference,na.rm = TRUE) %>%
          round(2),Mean = mean(Difference, na.rm = TRUE) %>%
          round(3),SD = sd(Difference, na.rm = TRUE) %>%
          round(3),n = n())
price_difference_summary_category
```

Category <chr>	Min <dbl>	Q1 <dbl>	Medi... <dbl>	Q3 <dbl>	Max <dbl>	Mean <dbl>	SD <dbl>	n <int>
Baby,Health& Beauty	-1.50	0.00	0.00	0.59	9.24	0.760	1.871	36
Bakery	-2.00	-0.06	0.00	0.21	2.50	0.210	0.994	28
Drinks & Tobacco	-18.00	-0.07	0.70	3.00	9.54	0.233	6.275	30
Entertainment & International	-3.06	-0.21	0.00	0.00	0.99	-0.259	0.677	32
Freezer	-0.50	0.00	0.00	0.92	7.00	0.573	1.396	32
Fridge	-0.51	-0.15	-0.01	0.00	1.29	-0.008	0.311	29
Household & Pet	-3.50	-0.02	0.00	0.41	8.25	0.227	1.762	30

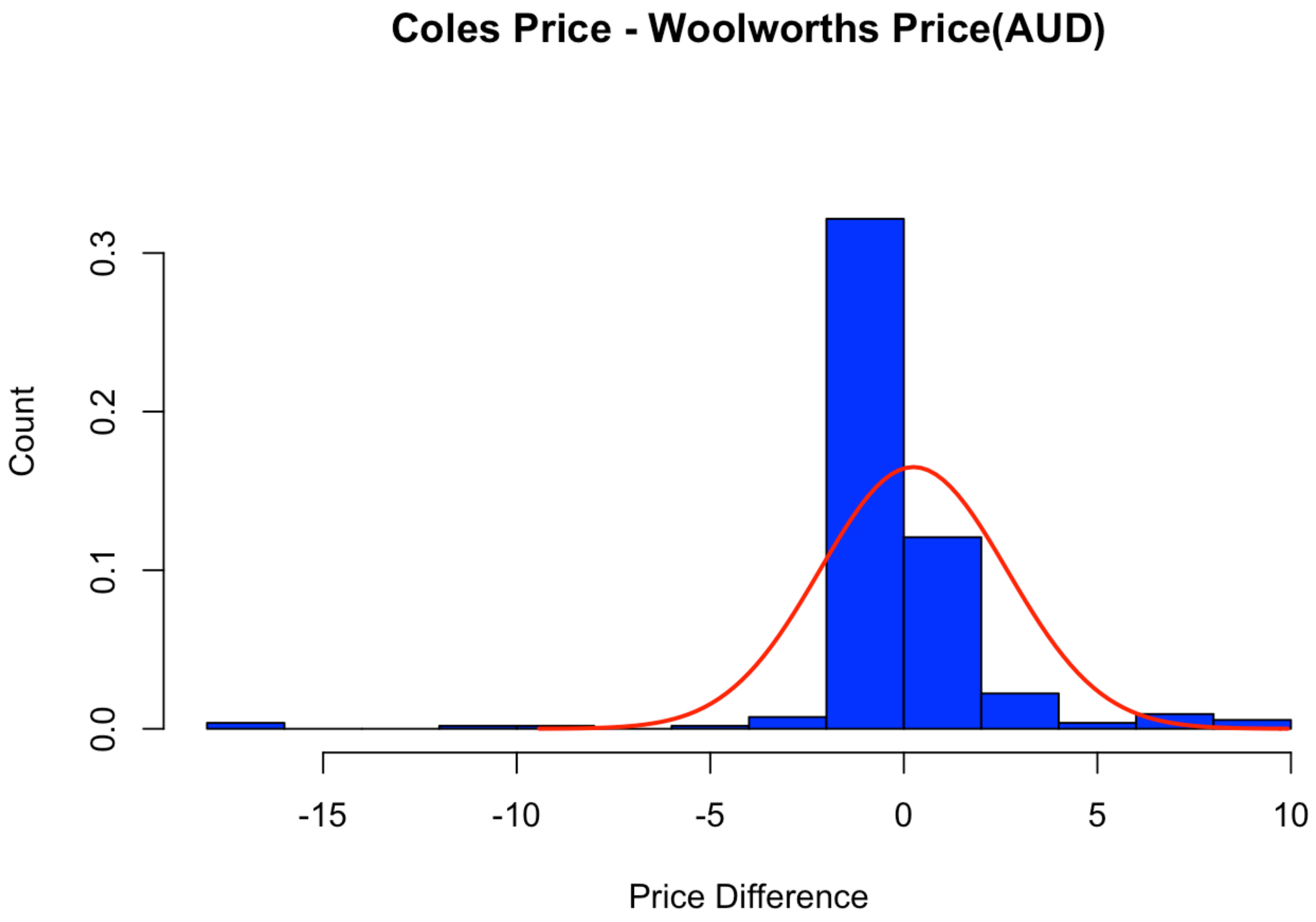
Meat & Seafood	-5.41	-0.30	-0.01	0.09	1.90	-0.215	1.581	17
Pantry	-1.20	0.00	0.00	0.73	6.00	0.393	1.084	35

9 rows

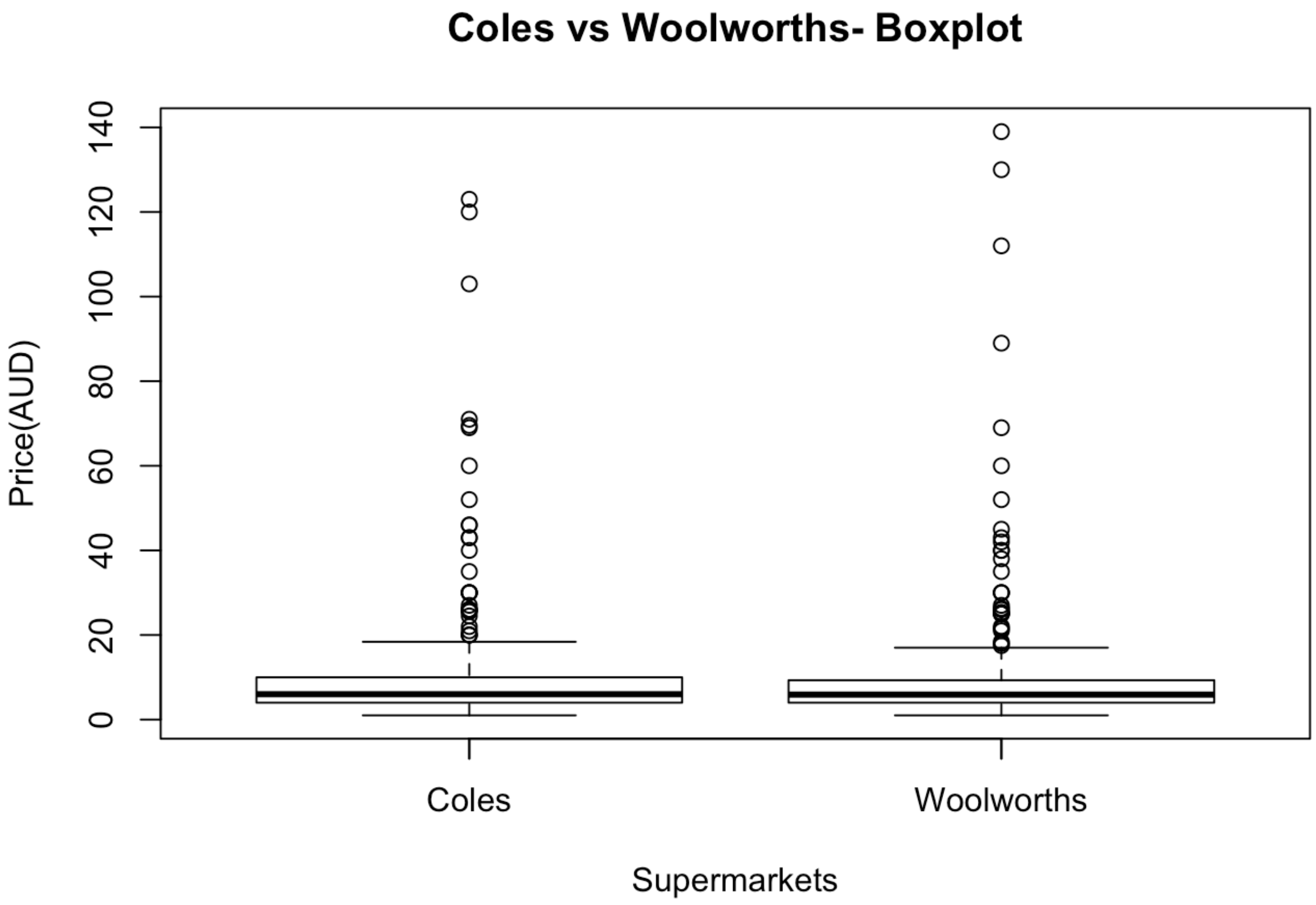
```
#Grouped Barplot to see the difference in PRices
ggplot(gathered_supermarts, aes(fill=Store, y=price, x=Category))+
geom_bar(position="dodge", stat="identity")+ggtitle("Coles Price Vs Woolworths Pri
ce")+ylab("Sum of Prices (AUD)")
```



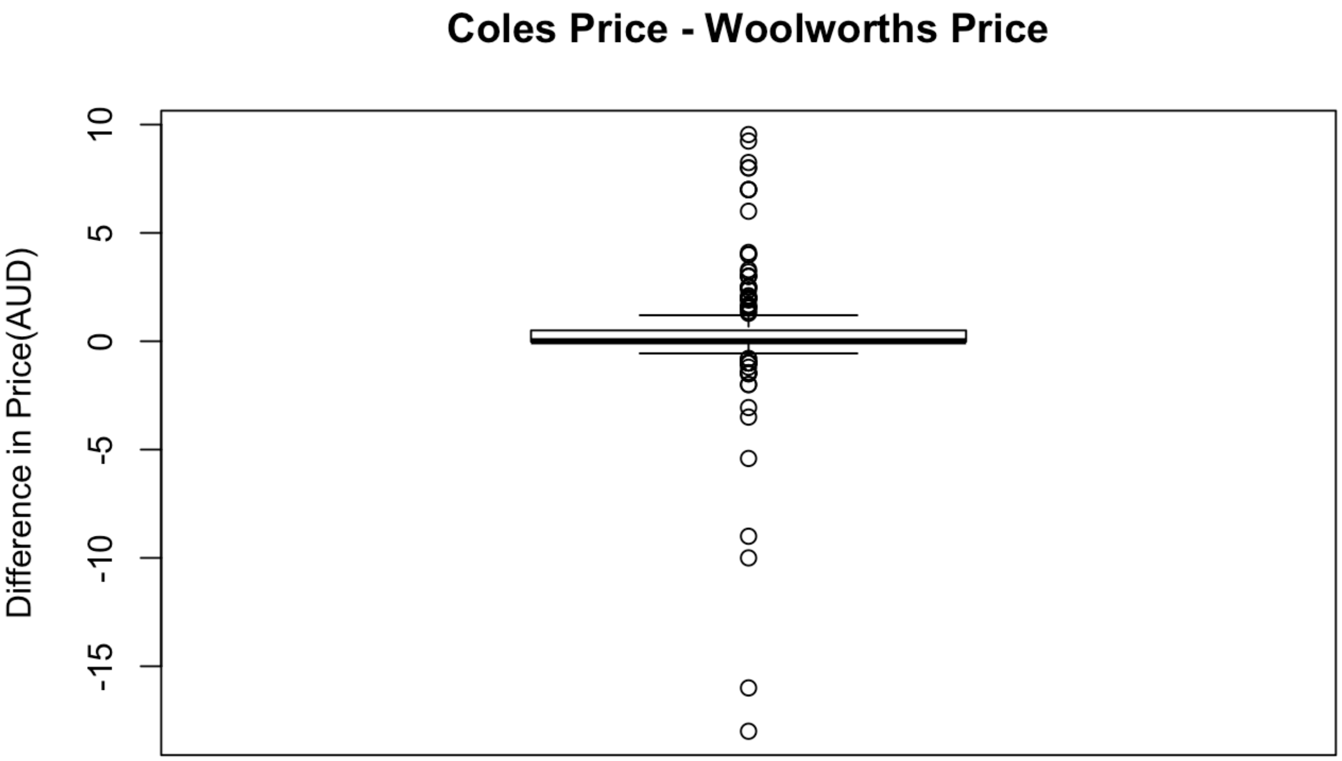
```
#Checking the difference in Prices Visually using Histogram
x <- seq(min(prices_data$Difference), max(prices_data$Difference))
mu <- mean(prices_data$Difference)
sd <- sd(prices_data$Difference)
prices_data$Difference %>% hist(col="blue", ylim=c(0,0.375), xlab="Price Differenc
e",ylab="Count",main="Coles Price - Woolworths Price(AUD)", prob = TRUE, breaks=10
)
curve(dnorm(x,mu,sd), xlim=c(mu-sd*4, mu+sd*4), col="red", add= TRUE, lwd =2)
```



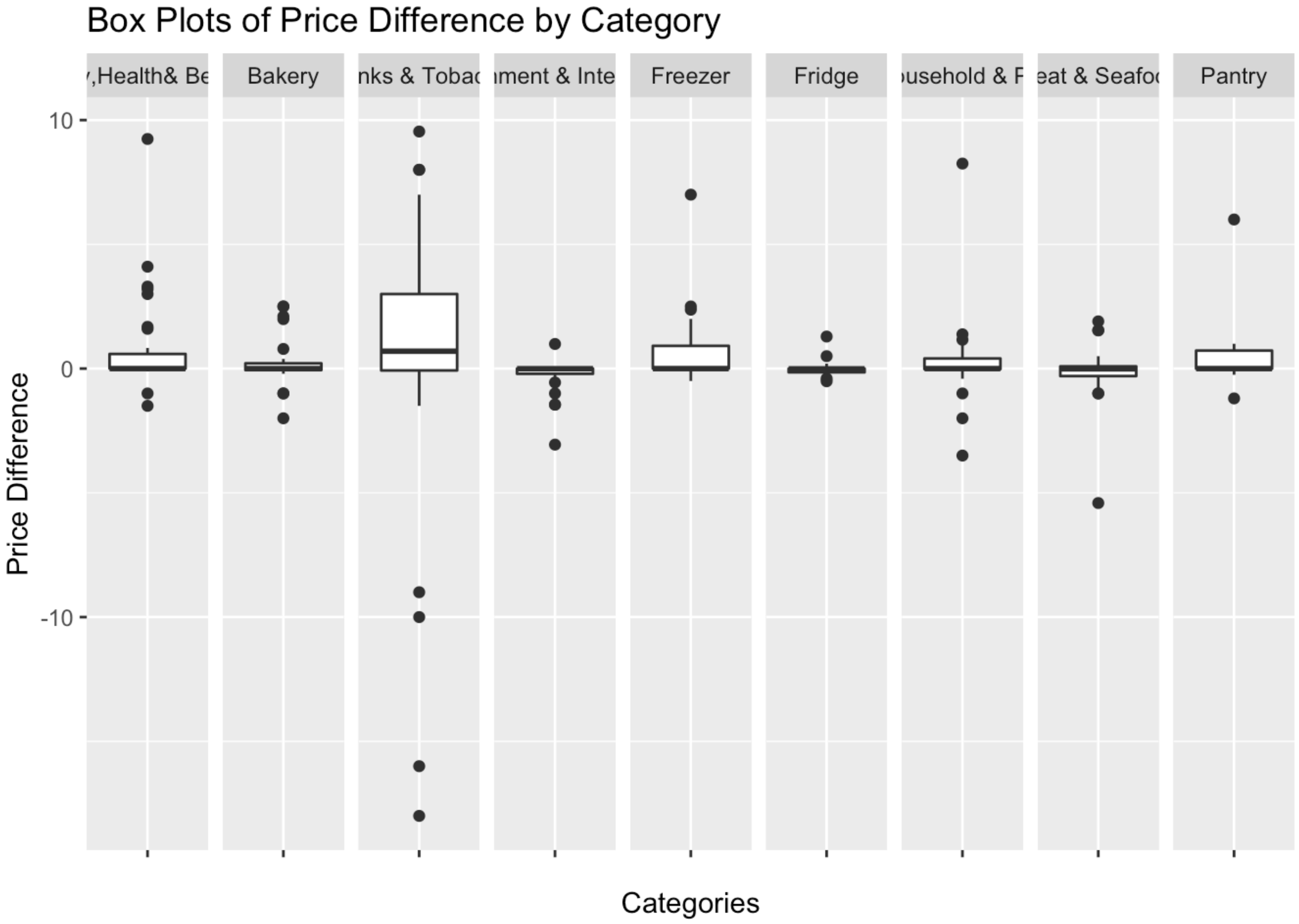
```
#Visualizing the price difference grouped by category, Further Analysis

boxplot(
  prices_data$Coles,
  prices_data$Woolworths,
  ylab = "Price(AUD)",
  xlab = "Supermarkets",
  main = "Coles vs Woolworths- Boxplot"
  )
axis(1, at = 1:2, labels = c( "Coles","Woolworths"))
```



```
prices_data$Difference %>% boxplot(main= "Coles Price - Woolworths Price",ylab="Di
fference in Price(AUD)")
```

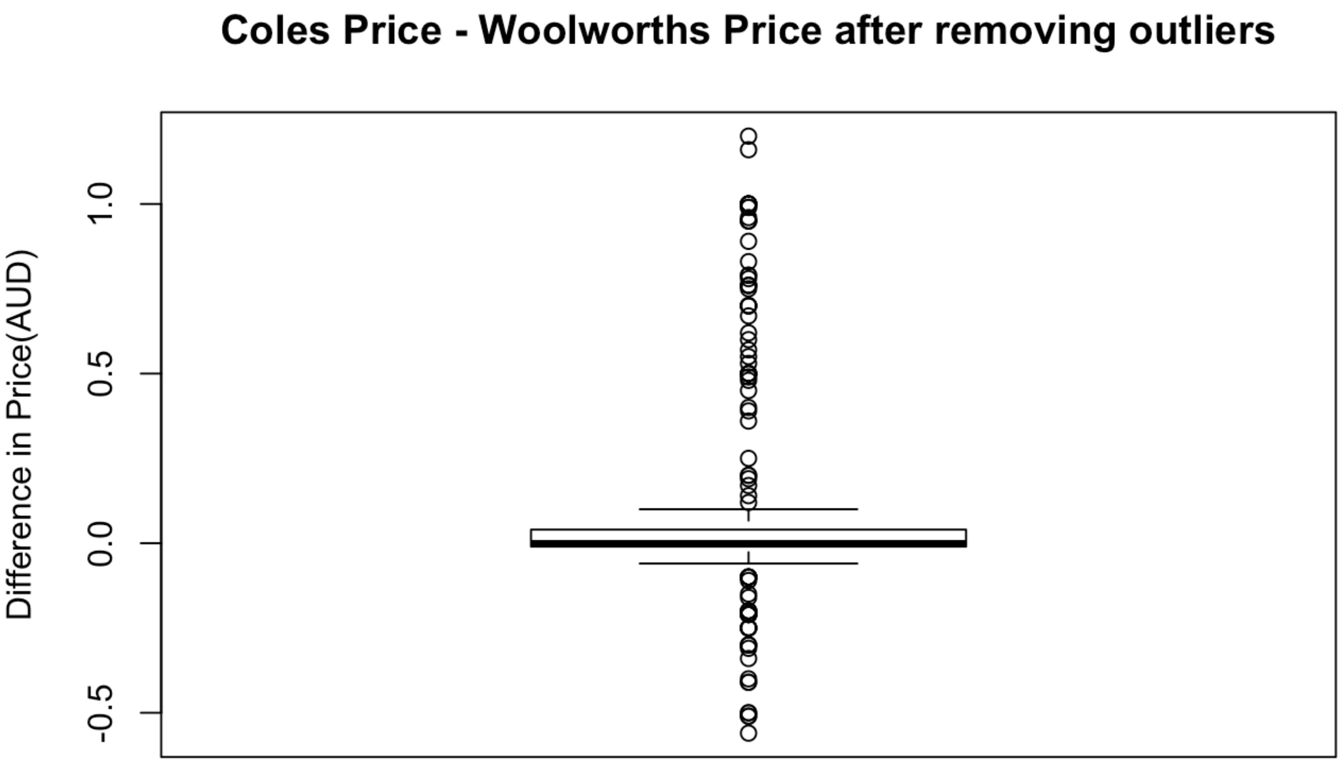


```
#Category wise boxplot

bp <- ggplot(prices_data, aes(x = "", y = Difference))
bp + geom_boxplot() +
facet_grid(.~Category) +
labs(title = "Box Plots of Price Difference by Category",
       y = "Price Difference",
       x = "Categories")
```

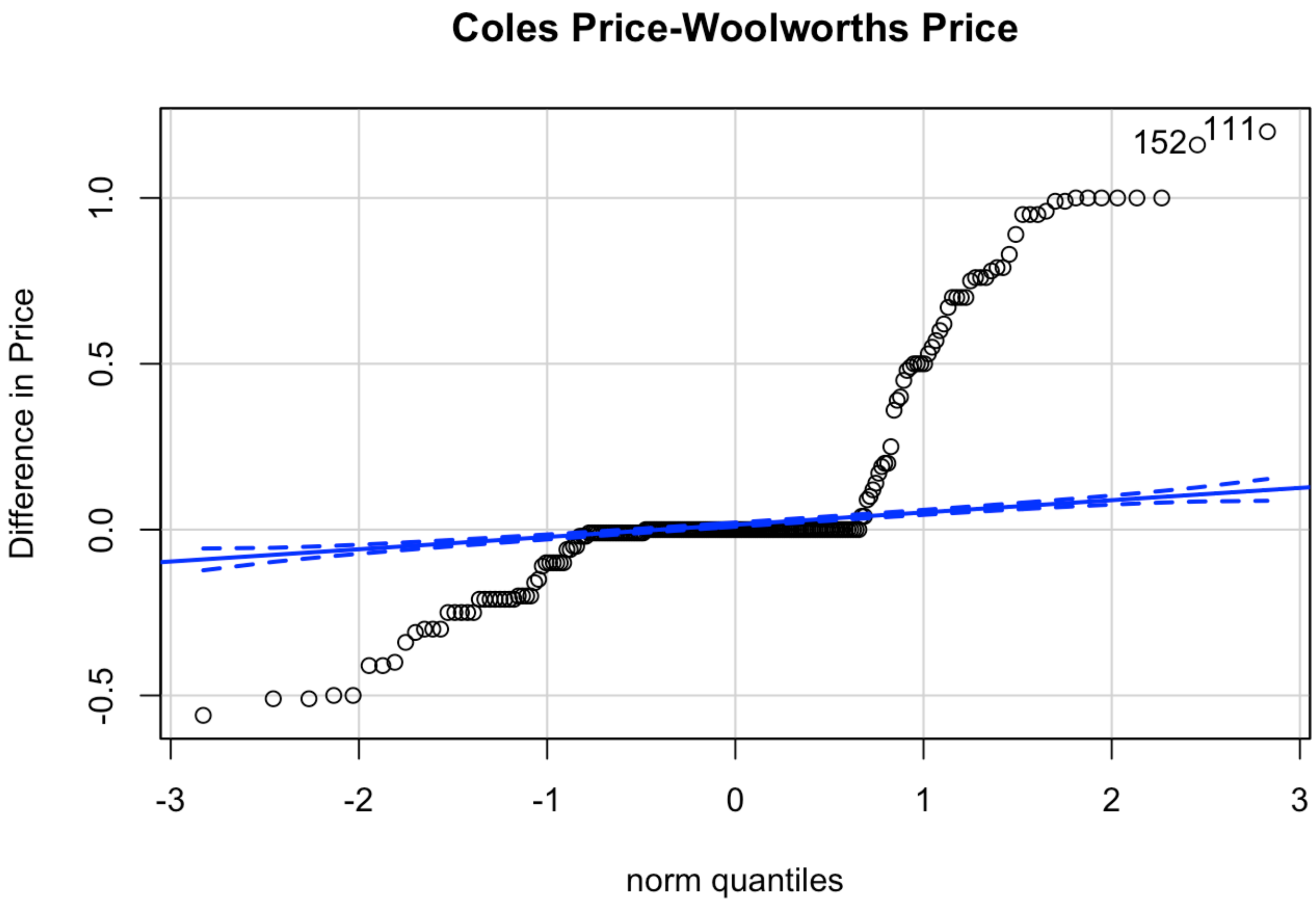


```
#Removing the outliers of the differences in prices

upper_fence <- 1.5*IQR(prices_data$Difference)+0.5

lower_fence <- -1.5*IQR(prices_data$Difference)-0.02

#Removing Outliers based on Category

upper_fence_bhb <- 1.5*(price_difference_summary_category$Q3[1] - price_difference_summary_category$Q1[1]) + price_difference_summary_category$Q3[1]

lower_fence_bhb <- -1.5*(price_difference_summary_category$Q3[1] - price_difference_summary_category$Q1[1]) + price_difference_summary_category$Q1[1]

upper_fence_bkry <- 1.5*(price_difference_summary_category$Q3[2] - price_difference_summary_category$Q1[2]) + price_difference_summary_category$Q3[2]

lower_fence_bkry <- -1.5*(price_difference_summary_category$Q3[2] - price_difference_summary_category$Q1[2]) + price_difference_summary_category$Q1[2]

upper_fence_dnt <- 1.5*(price_difference_summary_category$Q3[3] - price_difference_summary_category$Q1[3]) + price_difference_summary_category$Q3[3]
```

```
lower_fence_dnt <- -1.5*(price_difference_summary_category$Q3[3] - price_difference_summary_category$Q1[3])+ price_difference_summary_category$Q1[3]

upper_fence_ent <- 1.5*(price_difference_summary_category$Q3[4] - price_difference_summary_category$Q1[4]) + price_difference_summary_category$Q3[4]

lower_fence_ent <- -1.5*(price_difference_summary_category$Q3[4] - price_difference_summary_category$Q1[4]) + price_difference_summary_category$Q1[4]

upper_fence_frz <- 1.5*(price_difference_summary_category$Q3[5] - price_difference_summary_category$Q1[5]) + price_difference_summary_category$Q3[5]

lower_fence_frz <- -1.5*(price_difference_summary_category$Q3[5] - price_difference_summary_category$Q1[5]) + price_difference_summary_category$Q1[5]

upper_fence_frg <- 1.5*(price_difference_summary_category$Q3[6] - price_difference_summary_category$Q1[6]) + price_difference_summary_category$Q3[6]

lower_fence_frg <- -1.5*(price_difference_summary_category$Q3[6] - price_difference_summary_category$Q1[6]) + price_difference_summary_category$Q1[6]

upper_fence_hnp <- 1.5*(price_difference_summary_category$Q3[7] - price_difference_summary_category$Q1[7]) + price_difference_summary_category$Q3[7]

lower_fence_hnp <- -1.5*(price_difference_summary_category$Q3[7] - price_difference_summary_category$Q1[7]) + price_difference_summary_category$Q1[7]

upper_fence_mns <- 1.5*(price_difference_summary_category$Q3[8] - price_difference_summary_category$Q1[8]) + price_difference_summary_category$Q3[8]

lower_fence_mns <- -1.5*(price_difference_summary_category$Q3[8] - price_difference_summary_category$Q1[8]) + price_difference_summary_category$Q1[8]

upper_fence_ptry <- 1.5*(price_difference_summary_category$Q3[9] - price_difference_summary_category$Q1[9]) + price_difference_summary_category$Q3[9]

lower_fence_ptry <- -1.5*(price_difference_summary_category$Q3[9] - price_difference_summary_category$Q1[9]) + price_difference_summary_category$Q1[9]

#Tidying the Data (removing outliers)

price_data_tidy <- prices_data %>% filter(Difference < upper_fence &
                                          Difference > lower_fence)

gathered_supermarts_tidy <- gathered_supermarts %>% filter(Difference < upper_fence & Difference > lower_fence)

#Boxplot after removing outliers

price_data_tidy$Difference %>% boxplot(main= "Coles Price - Woolworths Price after
removing outliers",ylab="Difference in Price(AUD)")
```



```
#QQ-plot to check normality
price_data_tidy$Difference %>% qqPlot(distribution = "norm",main = "Coles Price-Wo
olworths Price",ylab = "Difference in Price")
```



```
## [1] 111 152
```

Hypothesis Test

- Hypothesis test has been conducted to check whether the mean difference of prices is significant enough to comment if Woolworths is cheaper than Coles.

NULL HYPOTHESIS:

- Mean of Coles and Woolworths price are equal.

```
μΔ= mu1 - mu2

H0:μΔ=0
```

ALTERNATE HYPOTHESIS:

- Mean of Coles and Woolworths price are not equal.

```
μΔ= mu1 - mu2

HA:μΔ≠0
```

This hypothesis would be tested using a paired t-test.

Assumptions- • Assumption of normality to be met for conducting t-test.

Normality- • As the sample size (269) is very large compared to the minimum sample of 30,normality of data can safely be assumed as per the Central Limit Theorem.

```
#General T-Test

gathered_supermarts <- gathered_supermarts %>%
filter( Difference < upper_fence & Difference > lower_fence)
price_t_test <- t.test(price ~ Store,
                       data = gathered_supermarts,
                       paired = T,conf.level = .95,
                       alternative = "two.sided")

#Removing the ouliers from each category

gathered_supermarts_bhb <-  gathered_supermarts %>%
                            filter(Category == "Baby,Health& Beauty" & Difference
< upper_fence_bhb & Difference > lower_fence_bhb)
gathered_supermarts_bkry <-  gathered_supermarts %>%
                             filter(Category == "Bakery"
& Difference < upper_fence_bkry & Difference > lower_fence_bkry)
gathered_supermarts_dnt <-  gathered_supermarts %>%
                            filter(Category == "Drinks & Tobacco"
& Difference < upper_fence_dnt & Difference > lower_fence_dnt)
gathered_supermarts_ent <-  gathered_supermarts %>%
                            filter(Category == "Entertainment & International" & D
ifference < upper_fence_ent & Difference > lower_fence_ent)
gathered_supermarts_frz <-  gathered_supermarts %>%
                            filter(Category == "Freezer"
                                   & Difference < upper_fence_frz & Difference > l
ower_fence_frz)
gathered_supermarts_frg <-  gathered_supermarts %>%
                            filter(Category == "Fridge" &  Difference < upper_fenc
e_frg & Difference > lower_fence_frg)
gathered_supermarts_hnp <-  gathered_supermarts %>%
                            filter(Category == "Household & Pet" & Difference < up
per_fence_hnp & Difference > lower_fence_hnp)
gathered_supermarts_mns <-  gathered_supermarts %>%
                            filter(Category == "Meat & Seafood" & Difference < upp
er_fence_mns & Difference > lower_fence_mns)
gathered_supermarts_ptry <-  gathered_supermarts %>%
                            filter(Category == "Pantry"& Difference < upper_fence_
ptry & Difference > lower_fence_ptry)

#T-test for each category

t.test(price ~ Store,
                       data = gathered_supermarts_bhb,
                       paired = T,conf.level = .95,
                       alternative = "two.sided")
```

```
##
##  Paired t-test
##
## data:  price by Store
## t = 2.6536, df = 26, p-value = 0.0134
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03130374 0.24647404
## sample estimates:
## mean of the differences
##               0.1388889
```

```
t.test(price ~ Store,
                       data = gathered_supermarts_bkry,
                       paired = T,conf.level = .95,
                       alternative = "two.sided")
```

```
##
##  Paired t-test
##
## data:  price by Store
## t = -0.014845, df = 19, p-value = 0.9883
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07099672  0.06999672
## sample estimates:
## mean of the differences
##                  -5e-04
```

```
t.test(price ~ Store,
                       data = gathered_supermarts_dnt,
                       paired = T,conf.level = .95,
                       alternative = "two.sided")
```

```
##
##  Paired t-test
##
## data:  price by Store
## t = 2.7746, df = 12, p-value = 0.01682
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.07846041 0.65230882
## sample estimates:
## mean of the differences
##               0.3653846
```

```
t.test(price ~ Store,
                       data = gathered_supermarts_ent,
                       paired = T,conf.level = .95,
                       alternative = "two.sided")
```

```
##
##  Paired t-test
##
## data:  price by Store
## t = -2.7654, df = 25, p-value = 0.01053
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.11810566 -0.01727896
## sample estimates:
## mean of the differences
##             -0.06769231
```

```
t.test(price ~ Store,
                       data = gathered_supermarts_frz,
                       paired = T,conf.level = .95,
                       alternative = "two.sided")
```

```
##
##  Paired t-test
##
## data:  price by Store
## t = 1.8968, df = 27, p-value = 0.0686
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01304919  0.33233491
## sample estimates:
## mean of the differences
##               0.1596429
```

```
t.test(price ~ Store,
                       data = gathered_supermarts_frg,
                       paired = T,conf.level = .95,
                       alternative = "two.sided")
```

```
##
##  Paired t-test
##
## data:  price by Store
## t = -1.8802, df = 24, p-value = 0.07228
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.09229993  0.00429993
## sample estimates:
## mean of the differences
##                  -0.044
```

```
t.test(price ~ Store,
                       data = gathered_supermarts_hnp,
                       paired = T,conf.level = .95,
                       alternative = "two.sided")
```

```
##
##  Paired t-test
##
## data:  price by Store
## t = 1.4762, df = 23, p-value = 0.1534
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04213752  0.25213752
## sample estimates:
## mean of the differences
##                   0.105
```

```
t.test(price ~ Store,
                       data = gathered_supermarts_mns,
                       paired = T,conf.level = .95,
                       alternative = "two.sided")
```

```
##
##  Paired t-test
##
## data:  price by Store
## t = -0.56601, df = 9, p-value = 0.5852
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2098609  0.1258609
## sample estimates:
## mean of the differences
##                  -0.042
```

```
t.test(price ~ Store,
                       data = gathered_supermarts_ptry,
                       paired = T,conf.level = .95,
                       alternative = "two.sided")
```

```
##
##  Paired t-test
##
## data:  price by Store
## t = 3.7519, df = 32, p-value = 0.0006993
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1239694 0.4184548
## sample estimates:
## mean of the differences
##               0.2712121
```

Interpretation

• GENERAL TEST: p < 0.001 alpha = 0.05 Confidence Interval = 95% As p < alpha, there is statistically significant difference between the mean prices of Coles and Woolworths. Therefore, we can safely reject the null hypothesis.

• Test conducted categorically: alpha = 0.05 Confidence Interval = 95%

• Interpretation of Null Hypothesis with respect to categories:

Baby,Health and Beauty,Drinks and Tobacco,Entertainment and International,Pantry reject null hypothesis where as Bakery,Freezer,Fridge,Household and Pets,Meat and Seafood fail to reject Null Hypothesis.

Discussion

• The findings from our analysis helped us in arriving at a conclusion that statiscally,the price of products in Coles is significantly higher than that of Woolworths .

• The average price of products in Coles is approximately $2.42 more than that of Woolworths.

• Nevertheless,a few categories like Bakery, Freezer, Fridge, Household and Pets, Meat and Seafood the mean prices in both stores is equal.

• The strenghth of our analysis is that category wise hypothesis test has been conducted in order to give a robust understanding.

• The limitation is that we wish to improve on increasing sample size per category so that the anaylsis would have been even more accurate.

• In future, we would like to consider specials and offers to be able to make a better judgement in terms of business perspective of both the stores.